Tomi Kokkonen

Evolving in Groups

Individualism and Holism in Evolutionary Explanation of Human Social Behaviour

DOCTORAL DISSERTATION

to be presented for public discussion with the permission of the Faculty of Arts of the University of Helsinki, in lecture room 107, Athena (Siltavuorenpenger 3 A), on the 10th of September 2021 at 13 o'clock.

Filosofisia tutkimuksia Helsingin yliopistosta
Filosofiska Studier från Helsingfors universitet
Philosophical Studies from the University of Helsinki

Publishers:
Theoretical Philosophy
Philosophy (Swedish)
Social and Moral Philosophy

P.O. Box 24 (Unioninkatu 40 A)
00014 University of Helsinki
Finland

Editors:
Samuli Reijula
Michiru Nagatsu
Thomas Wallgren

ISBN 978-951-51-7460-4 (paperback)
ISBN 978-951-51-7461-1 (PDF)
ISSN 1458-8331 (series)
Unigrafia, Helsinki 2021

Abstract

The main topic of this dissertation is to clarify the relationship between groups and individuals in the evolutionary explanation of human social behaviour. To the extent that we can understand human social behavioural traits as adaptations, are they adaptations of individuals (explanatory individualism) or groups (explanatory holism)? I will distinguish three different causal dimensions in an evolutionary explanation: proximate, developmental, and evolutionary proper. An evolutionary explanation makes (implicit) assumptions about how the behaviour is produced (the proximate dimension), and how it is reproduced (the developmental dimension). Both can involve either individual causal factors only or include supra-individual social factors. The main issue in the evolutionary dimension is whether group selection is an important factor in evolution. The group selection controversy is not, however, only about the nature of selection in the hierarchical biological organization, but also about the two other dimensions, as I will argue.

The first topic that I discuss is what evolutionary explanation of a behavioural trait is. I will develop an evolutionary functionalist account of how to individuate a behavioural trait, and I will discuss adaptationism in this context. I will show that there are three distinct defensible ideas of what an evolutionary function of a trait is, all of which are relevant. I will also distinguish between psychological, agentive, and behavioural traits in the human context, arguing that our usual way of classifying behaviour into traits is biased by folk psychology. After this, I will demonstrate how behavioural traits may be interactive traits that emerge in interaction and are not reducible to individual traits. This entails a non-individualist approach to adaptations even without group selection. As for the developmental dimension, I will discuss culture and innateness within the Extended Synthesis interpretation of evolution. I will discuss the complexities in understanding the roles of culture in evolution, and how exactly culture

contributes to holism. My main interest in this context is, however, innateness and nativist evolutionary psychology. I will develop a new definition for innateness and defend nativism, understood in this sense, as a plausible methodological choice, while at the same time I will highlight reasons for holistic alternatives of evolutionary psychology. Finally, I will discuss the group selection controversy. I will clarify some of the confusions in the debate using my work in the previous chapters about the other dimensions and their relevance to selection. In particular, I will clarify the difference between kin selection and group selection, and the controversy over group adaptations.

Key words: evolutionary explanation, social evolution, multilevel selection, group selection, culture in evolution, innateness, altruism, evolutionary psychology, evolutionary anthropology, folk psychology, mechanistic explanation

Table of Contents

Acknowledgments

The road to this dissertation has been long and winding, with detours and sidesteps to topics that have seemed relevant or just more interesting at the time. The long road has not been lonely, however. My thinking about the topics of this dissertation, as well as my philosophical thought in general, has evolved in interaction with a community, participating in research groups, reading groups, seminars, and societies, and other ways in which academic work is socially structured. I owe gratitude to many members of the surrounding academic community for support, feedback, inspiration, and making this journey mostly enjoyable.

I am most indebted to my supervisor Professor Petri Ylikoski. His lectures on philosophy of science, as well as willingness to engage discussions with an ignorant student played a pivotal role in my choosing philosophy of science as my speciality. His guidance and example played a pivotal role in making the first steps of my academic career possible. I am especially appreciative of his practical approach to supervising through collaboration, both in research and teaching, at those early stages, as well as all the advice and constructive criticism ever since.

I am also extremely grateful to my supervisor Professor Matti Sintonen. He has been not only an inspiring teacher and a role model as a philosopher of science, but he also provided invaluable advice on all things academic from research to more mundane practicalities while I worked in his research group, first as a research assistant, then as a researcher. His continuous support for this project has been crucial to making it possible.

A special thank you goes to Professor Uskali Mäki. While not a supervisor or an expert on most of the issues of this dissertation, the intellectual interactions with him have been an important contribution to my development as a philosopher. He is also largely responsible for bringing together the research community that has been crucial to my work.

Another special thank you goes to Professor Panu Raatikainen. The many discussions in informal settings that we had while I was a student must have taught me more about philosophy (both in content and form)

than the lectures I took, and he has been supportive in all my endeavours during my doctorate project. I consider him one of my mentors.

Most teachers whose lectures I have attended and with whom I have interacted have probably had some imprint in my philosophical processing. However, in addition to the above-mentioned, I want to single out Professor Gabriel Sandu who has been a much-needed example of a perceptive and deep thinker who is not a philosopher of science, to keep paradigmatic biases in balance. I would also like to thank him for his (partly successful) attempts to keep me focused on my dissertation when my interests were drawn elsewhere.

I have a privilege to have known some of the younger Professors since we were students together, and I have even a greater privilege to continue the philosophical discussions with them even now. The foundation of my philosophical thinking was largely developed through discussions with Jaakko Kuorikoski from undergraduate years on, and he has influenced this dissertation both through continuing discussions and his professional work. Antti Kauppinen and Teemu Toppinen have exceptional analytic minds that have always forced to increase the level of argumentation and depth of thinking.

If the memory serves, I started my philosophical argument with Hanne Appelqvist the first official day of my studies and it is still ongoing. The differences in philosophical paradigm have never been an obstacle to fruitful discussion, and the regular lunches with her have had a great significance both philosophically and personally over the years, for which I owe deep gratitude.

I cannot emphasize enough the importance of working as a part of a community. TINT Centre for Philosophy of Social Science, and the Helsinki Philosophy of Science community around it, has been a stimulating research environment and has provided important insights and feedback, even when the research of the group has mostly been unrelated to my dissertation. I would like to collectively thank all the people who have been members of this community over the years. In addition to people already mentioned, Sonja Amadae, Marion Godman, Till Grüne-Yanoff, Säde Hormio, Tero Ijäs, Tuukka Kaidesoja, Tarja Knuuttila, Rami

Koskinen, Aki Lehtinen, Caterina Marchionni, Carlo Martini, Michiru Nagatsu, Jani Raerinne, Kristina Rolin, Anna-Mari Rusanen, Mikko Salmela, Päivi Seppälä, Tuomas Vesterinen, Julie Zahle, and many others (who my memory fails to conjure up, and I apologize for that) have provided insightful feedback and discussions. Emrah Aydinonat, Ilmari Hirvonen, Mika Kiikeri, Inkeri Koskinen, Magdalena Małecka, and Samuli Reijula have gone beyond commenting to co-authoring talks and papers, although not always on the topic of the dissertation. I would especially like to thank Inkeri. My detour into the philosophy of humanities with her was probably not productive for the dissertation project, but indubitably made me a better philosopher. I must also give special thanks to Judith Favereau and Luis Mireles Flores for exceptional colleagueship. Alkistis Elliott-Graves was among the first people I got to know from the international philosophy of biology community – and little did I know when we first met in Australia that she would eventually become a close colleague in Helsinki and play such an important role at the last stages of my dissertation project.

Last but not least of the Helsinki Philosophy of Science community that I want to thank are Pekka Mäkelä and Raul Hakli. Not only have they provided insightful feedback to some of my work, but they also co-lead the research group that I am currently working in. The group has already proven to be an inspiring and productive research environment, and I am eagerly looking forward to continuing working with my brilliant colleagues Dina Babushkina, Dane Leigh Gogoshin, Olli Niinivaara, and Pii Telakivi. I would also like to thank Pekka for all his help and all the discussions (philosophical or otherwise) since I started working as a research assistant. Discussing philosophy with such a clear and precise thinker has been both pleasurable and instructive.

I am also grateful to the international philosophy of biology community, which has been very welcoming and helpful, and provided important feedback throughout this process. I would like to thank especially Patricia Churchland, Ellen Clarke, Carl Craver, Paul Griffiths, Philip Kitcher, Beate Krickel, Stefan Linqvist, Robert Richardson, Richard Samuels, Elliot Sober, and Michael Weisberg for useful discussions on various

topics in this dissertation. Special thanks go to Shereen Chang for the numerous in-depth discussions around the world. Particular thanks are due to Maria Kronfeldner and Grant Ramsey who kindly agreed to read the whole manuscript as preliminary examiners and whose suggestions and encouraging comments helped to finish the work.

I would also like to thank Johanna Ahola-Launonen, Jussi Backman, Ferdinand Garoff, Kaisa Heinlahti, Tarna Kannisto, Markku Keinänen, Yiannis Kokosalakis, Anssi Korhonen, Simo Kyllönen, Kaisa Kärki, Arto Laitinen, Vili Lähteenmäki, Lilian O'Brien, Anna Ovaska, Ville Paukkonen, Tuomas Pernu, Ilkka Pättiniemi, Markku Roinila, Ninni Suni, Tuukka Tanninen, Sanna Tirkkonen, Pilvi Toppinen, Leena Tulkki, Simo Vehmas, Jaana Virta, Anita Välikangas, and everyone else who I should mention but fail to do so, as the wider academic community that has contributed to this process one way or another. I am also grateful to the philosophy administrative staff, Ilpo Halonen, Auli Kaipainen, Terhi Kiiskinen, Karolina Kokko-Uusitalo, and Tuula Pietilä for all their help over the years.

Special thanks to Johanna Sinkkonen and Janne Tompuri (and their children Aarni and Niila), and Sanna Nyqvist (and her family Jyrki Hakapää and Meri) for lasting and meaningful friendship since the early student days.

Of all the groups we are part of, family is of special importance. I would like to thank my mother Aili Kokkonen and my sisters Tarja Kokkonen and Terhi Rinnemäki (and her family Jari, Leevi, Luukas and Lauri), as well as my in-laws Saara and Eero Kaakinen, Kaisa Kaakinen, and Timo Kaakinen (and his family Henna Jaurila, Maiju and Saku), for showing why this is so.

My final thank you goes to Leena Kaakinen, for love, companionship, patience with this project, and for everything that ultimately matters.

Helsinki, 15th August 2021
Tomi Kokkonen

1. Introduction

Humans are social beings who evolved in groups. If we take an evolutionary perspective on human social behaviour, should we understand its evolutionary functionality from the individual or group perspective? In other words: if we apply adaptationist heuristics in our attempts to understand human sociality, should we adopt individualist or holist approaches and frameworks? This is a question about the methodological choices in evolutionary explanations for social behaviour. The main aims of this dissertation are: 1) To explicate the differences between individualist and holistic explanations (and individualistic and holistic presuppositions in explanations) in the context of the evolution of human social behaviour. This involves three *explanatory dimensions* in which the relationship between individual and supra-individual causal factors may matter: proximate, development, and evolutionary. 2) To make arguments for explanatory holism. As the first approximation, the individualist alternatives approach human social behavioural adaptations as individual adaptations in a social context, while the holistic alternatives approach humans as inherently social beings, and the evolutionary functionality of social behaviour is an approach from the group perspective. I call these approaches *explanatory individualism and holism* regarding *evolutionary social science.*

The central theme in this issue has been the controversy about the levels of selection, but in the human context, the relevance of the group structure is connected to other questions. What is an adequate way to identify the behavioural traits and the mechanisms that produce social behaviour? What is the role of socio-cultural environment in the development? How are these issues connected? In other words, what is the relevance of *supra-individual level causal factors in various explanatory dimensions* for the *evolutionary explanations of human social behaviour*? To answer these questions, I will dissect some aspects of proximate, developmental, and evolutionary explanations of social behaviour, and their connections, from a biological point of view. The

distinction between these explanatory dimensions is based on Niko Tinbergen's (1963) classic distinction between the four questions of behavioural biology. I will also connect Elisabeth Lloyd's (1992, 2001 & 2017) discussion on the different issues within the group selection controversy to these dimensions. I will concentrate especially on the proximate dimension and argue for a form of holism about social behavioural traits. To prepare to address these questions, I will begin by discussing evolutionary explanations of behavioural traits in general. I will discuss this through a combination of the *contrastive-contrastive theory of causal explanation*, the accompanying *manipulationist account of causation*, and the *New Mechanistic Philosophy*.

The main questions of the thesis fall in the intersection of philosophy of biology and philosophy of human behavioural sciences.[1] The topic has direct relevance for some areas of empirical research as well; in particular, there are consequences for methodological choices in human evolutionary sciences. Many of the points emerging along the thesis will be implicitly critical of some common approaches in evolutionary human sciences, but the aim is to be constructive and argue for better approaches instead. I will not, for example, use space to criticize evolutionary psychology. Instead, I will articulate some consequences for evolutionary human sciences from my discussion. My discussion will also have wider philosophical consequences for theory of explanation, how we understand human behaviour and action, the underpinnings of human sociality and morality, as well as "human

[1] "Human behavioural sciences" refers to the multidisciplinary set of fields that aims to understand behavioural interactions between humans and their environment, and includes various research fields in psychology, anthropology, and biology, and some unique fields. The philosophical issues related to human behavioural sciences are customarily included in the philosophy of social sciences, the philosophy of psychology, and the philosophy of biology, but this characterization is more accurate. The empirical fields in focus are evolutionary psychology (broadly understood), biological anthropology, and evolutionary biology proper, but I will discuss some issues related to various other parts of behavioural sciences, psychology, anthropology, and biology.

nature". The secondary field of philosophy that the thesis belongs to could be called *naturalistic philosophical anthropology:* a philosophical analysis of (one part of) the scientific image of human being and the philosophical consequences of this image.[2]

I will now briefly discuss the concepts of levels and explanatory dimensions, and what I mean by individualism and holism having three explanatory dimensions in evolutionary explanation. After this, I will make a few remarks on evolutionary explanation in the human context, this being a contested issue. I will finish the introduction by giving a more precise characterization of the research question and an outline of the contents.

1.1. Levels and Dimensions

The topic of the dissertation is about how the various *levels of biological organisation* are related to each other in three different *dimensions of explanation* (proximate, developmental, and evolutionary), and how these dimensions are related to each other within the evolutionary framework. The motivation behind this endeavour is to clarify the methodological basis of evolutionary social science: when and how is the group level relevant to the explanatory point of view? The main substance of the dissertation will be the explication of the difference between individualistic and holistic approaches on these explanatory dimensions and how these issues are connected to each other in the substantial presuppositions of evolutionary explanation. The facts about other dimensions – proximate and developmental – have

[2] Peter Godfrey-Smith (2009) has re-introduced the old concept of *philosophy of nature* to distinguish the philosophical work on the subject matter of science from the philosophy of science focusing on methodology, epistemic practices, and related epistemological questions. "Philosophical anthropology" is a more fitting expression when these issues concern humans, but given its other uses, the attribute "naturalistic" is apposite.

indirect consequences for the debate on the levels of selection in the human context. These dimensions are related to independent explanatory questions, but the answers and their explanatory presuppositions are connected. Before articulating the main idea in more detail, I will clarify what I mean by "levels" and "dimensions".

1.1.1. The Concepts of Level and Explanatory Dimensions

There are different notions of level. The metaphor of "higher" and "lower" levels may refer to ontological fundamentality, degrees of abstraction, structural hierarchies, part-whole relationship (for example, in composition or in mechanistic explanation schemes), aggregation, or simply the scale (see Craver 2007 & 2015; Brooks 2019). The various levels, regardless of how they are characterized, are often considered to form quasi-separate layers with their own generalizable regularities. According to the traditional, "layer-cake model" of the organization of science, there is a hierarchy of "basic" sciences that study different levels and their unique characteristics, while the relationship between the levels can be studied as wholes (Oppenheim & Putnam 1958). This view seems to require that the ways in which levels are characterized capture more or less the same hierarchical organization of the reality. This does not seem to be the case, and, consequently, there have been calls to either stop using the misleading metaphor altogether (for example, Potochnik & McGill 2012; Eronen 2013 & 2015; Thalos 2013) or to use it in a pluralistic manner, recognizing multiple "dimensions" of hierarchy (for example, Wimsatt 1974, 1976 & 1997; Craver 2001, 2007 & 2015; Brooks 2019). Daniel S. Brooks, in turn, has recently argued that the notion of level should be understood as a tool for structuring problems (Brooks 2019). Levels-talk is a way to relate research questions and approaches that are distinct but systematically related. It is an imprecise but productive notion, and its content and relevance depend on the local epistemic goals. Following this pluralistic, pragmatic, and explanatory-goal dependent notion of levels, I consider "levels" to be a part of how we frame some questions in

philosophy and science. Furthermore, which dimensions of levels coincide in which cases is context-dependent. It is possible that the properties associated with levels cluster globally (see Wimsatt 1976; Eronen 2015; Brooks 2019) or locally. I will not make any specific assumptions about this. I will employ the concept of level in a narrow and local way, but in a couple of different senses.

The primary meaning of "level" in this context will be the *level of biological organization*. This notion is still somewhat imprecise as a general notion (Potochnik & McGill 2012; Eronen 2013) but the relevant levels of organization for the topic at hand are sub-individual processes, individuals, groups of individuals, and the population.[3] These are clear enough in what they are and how the levels form a mereological hierarchy. I do not assume any other level-relations to hold automatically. The causal and explanatory relationships between these levels (and the individual and group levels especially) are the subject matter of the dissertation. What I call "dimensions" (proximate, developmental, and evolutionary) are sometimes also called "levels" (see Sherman 1988; Mitchell 1992; Reeve & Sherman 1993; Longino 2013). This scheme is derivative of Niko Tinbergen's famous "Four Questions of Ethology": cause, development, evolution, and function (Tinbergen 1963). I will discuss functions and evolutionary functional explanations, as well as its relation to the "dimensions", at length later. They do not presuppose integrated fields of explanation; there are different, separate explanatory approaches within each. Neither are the approaches partial replies to one "big question" of a field. Moreover, explanations in different dimensions are not indifferent to facts of each other. They are, however, three different categories of

[3] It should be noted here that genes are not a level of biological organization. They are, of course, a part of the structure of the cell and they play a crucial role in evolutionary explanations, but their role is very different from the one that is captured by this particular way of thinking about biological hierarchy. They play no role in the perspective taken here. I will discuss genes later; for now, this can be taken as a stipulation. The "lower levels" of biological hierarchy (such as cells) are outside the scope of this thesis in any case.

research questions that should not be conflated. I will examine their relation in general and in the context of the case at hand in particular at some length later.

There are other relevant concepts of level. First, the *mechanistic levels* in the mechanistic philosophy of science. These levels do not always coincide with the organizational levels (see Carver 2001, 2007 & 2015). I will argue that mechanistic thinking connects the explanatory *dimensions*, but this is a substantial claim about the relationship between kinds of explanatory projects in a very local case (evolutionary explanation of behaviour), not a statement about biological explanation in general. Second, there are *levels of abstraction* (or *analysis*) in the context of psychological explanation (for example, Marr 1982), which, again, do not neatly coincide with either of the two other level-distinctions. Furthermore, there is the distinction between different *stances* (Dennett 1987) that is substantially different from all of the above, although it is related to the levels of abstraction in psychological explanations. In my discussion, the notion of level referred to will be clear from the context and all these issues will be discussed later. Now I will give a general characterization of levels of the organization and the explanatory dimensions and how they are related.

1.1.2. Individualism and Holism in Three Dimensions

Humans are social animals and adapted to their ancestral environments as groups. My main question is, should we understand this adaptation process from the individualistic or holistic (that is, social or group) perspective? What kind of entities are groups, when seen from an evolutionary functionalist perspective? The main evolutionary issue is about whether group selection exists and is an important enough factor to be accounted for, and whether there are group adaptations. I will discuss group selection and the levels of selection issue directly in the last chapter, and, as we shall see, the issue is messier than just a question about the levels of biological organization at which natural selection operates. As Elisabeth Lloyd (1992, 2001 &

2017) has argued, the levels of selection debate has four different components that are distinct issues, often confused in the debate: the *interactor question*, the *replicator question*, the *beneficiary question*, and the *manifestor-of-adaptation* question. I will return to this in detail later, but the rough idea is the following. The interactor question is about the levels on which the selection processes take place – this determines fitness consequences. The replication question is about the level at which the replication process takes place, which determines *whose* fitness we are talking about. The manifestor-of-adaptation question is about the level of organization at which the evolving traits exist. I will call these three questions individualism versus holism issues in the three causally relevant dimension in the evolutionary explanation: *evolutionary*, *developmental*, and *proximate*, respectively.[4] These causal explanatory dimensions are connected in the evolutionary explanation even without the levels question, and an evolutionary explanation makes assumptions about all of them. I will discuss their connection without the levels-aspect at length first, before discussing the individualism and holism issue in each dimension, and in connection to each other.

In the human context specifically, the proximate question is about whether there are only individual behavioural traits evolving, or also supra-individual traits that are of the group or of the interactions within the group, and whose evolutionary function cannot be understood at the individual level alone. I will argue that some social traits are the latter. I will argue further that although this does not necessarily require group selection, it can facilitate group selection. The developmental question is about whether the traits under selection are transmitted through individualistic processes (roughly speaking, they are innate) or if the development is systematically influenced

[4] The beneficiary question is about the level at which the entities that *ultimately* benefit from the evolution exist, which is not an explanatory question but a question about the ontology of evolution. I will not discuss this subject at length.

by the social environment in a way that binds the outcome to belonging to a particular group. Holism in this dimension has two consequences: it facilitates proximate holism by causing uniformity, and it facilitates group selection by connecting the fitness of individuals.

To illustrate what I mean by individualism and holism in this specific context, here are two examples. An example of purely individualistic evolutionary human science would be *nativist evolutionary psychology,* which has individual behavioural dispositions (identified with an individual's psychological states and/or mechanisms) as its *explananda*, assumes all interesting properties to be (mostly) innate, and employs individualistic selection models in its explanations. A purely holistic evolutionary human social science would consider the *explananda* to be the properties of groups that are assumed to be transferred (partly) culturally (and through other means of social transmission) and explained using group selection models. Various combinations of these three forms of holism are possible and there are more moderate and more extreme positions on each dimension. I will not defend or criticize any particular approach as such, for two reasons. First, the truth of the matter in each aspect is partly an empirical issue. Second, the correct approach may turn out to be different with respect to different types of social behaviour (when there is an adequate evolutionary approach in the first place, which is not always the case). It is possible that there will be no methodologically and theoretically unified evolutionary human science, but a pluralistic combination of different approaches instead. Approaches that seem to be problematic in some cases (and have been heavily criticized for good reasons, such as the nativist evolutionary psychology mentioned above) may still work in others. I will not discuss this topic in this dissertation directly. Instead, I will focus on explicating the differences in presuppositions that can be conceptualized as "individualist" and "non-individualist" alternatives, and the consequences of those presuppositions for evolutionary explanations, without taking sides. My dissertation will constitute a *conditional argument* for holistic approaches by explicating the criteria for the methodological choices regarding the issue. The

satisfaction of these conditions is an empirical and case-by-case question that cannot be given a general answer.

There is an old but ongoing debate in the philosophy of social sciences on whether social interactions can be adequately understood and explained in terms of individuals and their behavioural dispositions, or if a more holistic approach, building on "higher-level" properties**,** is needed (see Zahle & Collin 2014a; Ylikoski 2017). To put it simply, individualism is a position according to which individuals, their reactions to the environment (partly constituted by other people), and the aggregative consequences of those reactions are sufficient for understanding social phenomena.[5] According to holism, there are social structures that are not reducible to individuals and their dispositions alone, as well as cultural meanings and norms that play a role in explaining behaviour. The questions I am exploring are analogical in part and have some substantial overlaps with the issues in this debate. The proximate dimension is about the sufficiency of individual perspective to address what constitutes social behavioural traits in both cases, and the role of social learning and culture in the developmental dimension is an issue in both. Some evolutionary psychologists used to predict that evolutionary psychology would take

[5] To be more precise, there are two different but related areas in the individualism vs. holism debate in the philosophy of social sciences: the ontological status of social entities, and the methodological issues about to what extent should the research concentrate on individual and social level properties in e.g. explanation. On the methodological side, there are two different issues: the dispensability debate and the microfoundations debate. In the dispensability debate, the issue is about what the proper focus of explanation is: individuals (*methodological individualism*), social phenomena only (*strong methodological holism*), or both (*weak methodological holism*). What exactly counts as what position (e.g., what is the border between individualism and weak holism), is a contested issue itself. The microfoundations debate is about whether holistic causal claims need to be supplemented with mechanistic accounts that show how the social level causal relations emerge from individual-level interactions. (Zahle & Collin 2014b.)

over the social sciences as the naturalist foundation for understanding human sociality. From the social science point of view, this project would clearly have been an example of methodological individualism. Some of the responses from the social scientists included pointing out the holistic nature of human societies – but very similar objections could be made to evolutionary psychology as a form of evolutionary human social science already. Social sciences are, however, mostly about larger-scale social phenomena and kinds of interaction and social institutions that are too recent for these debates to be directly relevant to the topic. The possible direction of influence goes in the other direction: if there are good reasons to think that human sociality is fundamentally built on supra-individual connections that cannot be understood individualistically, that *could* become an argument for holism in social sciences – but that is a further issue that I do not go into. Additionally, there are some similarities in how to interpret seemingly individualist models (sometimes even the same game-theoretical models that travel between substantially distinct disciplines) and, as we will see, the issues about agency in understanding social behaviour are issues for evolutionary social science, too. These connections are, however, a side issue to the main topic and will not be discussed.

The third explanatory dimension I discuss is the group selection controversy within evolutionary biology (and in its philosophy). Can evolution of social behaviour be adequately described and explained in terms of the characteristics of individuals and their fitness differences alone (as individual adaptations to the environment partly constituted by other individuals), or is the evolution of behaviour sometimes due to its "higher-level" fitness consequences? This is the debate between *individualist* and *group-level* (or, rather, *multi-level*) *selection* theorists (see Okasha 2006). The levels of selection debate is vast and most of the issues are not relevant here, but I will discuss some issues in the last chapter of the dissertation.[6]

[6]The methodological Individualism issue in evolutionary biology is substantially distinct from the methodological individualism issue in social sciences,

In all three dimensions, the object of interest is social behaviour in the context of a wider social group. The issue is whether the perspective should be individualist, or the group more holistically. This makes the terminology of "individualism" and "holism" appropriate whether the connection to the similar issues within the philosophy of social science are substantial or merely analogical.

I will discuss all three explanatory dimensions only from the perspective of evolutionary explanation, which narrows down and defines the objects of the discussion in proximate and developmental dimensions. The *proximate dimension of evolutionary explanation* relates the behaviour being explained to its proximate causes. This is a central issue for individuation of the *explananda* of the evolutionary explanatory questions. Is the object of the evolutionary explanation the individual dispositions or the forms of interaction in which they participate? The main focus of the discussion will be on altruistic and prosocial behaviour, since this is the context that most clearly evokes the possible inadequacy of the individualistic approach. I will frame the issue in terms of *evolutionary functionalism*. In this framework, the function of a behavioural trait is what it does to increase the adaptivity of individuals or groups of individuals who participate in the behaviour, and this function individuates which occurrences of behaviour form a behavioural trait. The behaviour is a product of capacities and behavioural dispositions, and these capacities and dispositions constitute a mechanism for the behaviour. These parts may be properties of one or several individuals, which in turn determines whether we should consider behaviour to be individualistic or holistic. The approach is not meant to be a universal approach to behaviour, but I will

---

although some forms of structural functionalism might have appealed to some rudimentary evolutionary ideas. Elliot Sober (1980) and David Sloan Wilson (1989), however, have pointed out the parallels between the two and used the terminology of individualism and holism in the context of levels of selection, but in an obviously analogical way instead of connecting the debates as such.

argue that it is an adequate framework for thinking evolutionary explanations of some social behaviour.

Another dimension is the *developmental dimension of evolutionary explanation*. Cultural inheritance is an important alternative route for genetic inheritance for passing on behavioural traits. Should we understand this individualistically or through holistic cultural frameworks? Does culture, to some extent, "build" us to be part of a bigger whole, and if so, what is the evolutionary significance of this characteristic of our development? These are familiar questions, again, from the individualism and holism debate in the philosophy of social science, and they are partly empirical questions, but I will focus on their meaning and consequences for evolutionary explanation by discussing how to conceptualize the causal interaction between the individual and the developmental environment, including the social and cultural environment. The importance of this issue lies especially in the link between group selection and cultural evolution proposed by some evolutionary human scientists (for example, Boyd & Richerson 1985 & 2005; Wilson 2002), which is juxtaposed with the position of some evolutionary psychologists that the true locus of evolutionary explanations is the innate structure of the human mind. Much of the discussion will be on the concept of innateness and its explanatory relevance for evolutionary explanations, and on how culture figures as a significant route for inheritance in human context.

Lastly, I will discuss the consequences of these two dimensions for the relation between the individual level and supra-individual level in evolutionary explanations and to the levels of selection debate. I will discuss the levels of selection issue in a specific context only: humans. This particular context may be a special case and the results are not necessarily generalizable to, for example, *eusocial* insects like (most) ants and (some) bees, *presocial* mammals like wolves, chimpanzees, and meerkats, or any other forms of sociality found in non-human animals, no matter how analogical they seem to be with some aspects of human sociality. On the other hand, this context excludes

some issues that might be relevant in some other contexts of evolution of social behaviour.

1.2. On the Evolution of Human Social Behaviour

Behaviour is a more complicated object for evolutionary explanation than physical characteristics and social behaviour is trickier than behaviour in general. Social behaviour involves fitness effects for more than one individual at a time and the selective environment of the evolution of social behaviour is in continuous change, which causes the evolutionary functionality of the behavioural trait to be in constant flux. The evolution of social behaviour became an object of systematic study only after a new framework of mathematical models was introduced in the 1960s and 1970s, particularly by William D. Hamilton (1964a & 1964b), Robert Trivers (1971 & 1974), and Edward O. Wilson (1975), who gave the field its name, *sociobiology*, and systematized it (see also Dawkins 1976 and Alcock 2003). Soon after this, the same explanatory models were applied to human social behaviour, with varying degrees of success (see Kitcher 1985 for a review and criticism of much of the early work). After the partial failure of these early attempts, as well as some failures in the sociobiological approach overall, several different styles of evolutionary human sciences were developed, dealing with social behaviour that ranged from the general basis of the socially tuned mind and cooperation to specific issues concerning, for example, family relations, mating, religion, morality, and so on.[7]

There is a tendency in evolutionary human sciences to be methodologically individualist and this is partly for good reasons. The early forms of what was known as group selection (e.g. Wynne-Edwards 1962) had serious problems in its unsupported, ambivalent assumptions about the "good of the group" and could not come up with

[7] I will briefly review evolutionary human social sciences in chapter 4.

a mechanism for group-level selection to overcome individual selection (see Maynard Smith 1964; Williams 1966). The biggest problem was that there is always individual selection between the members of a group and therefore, if all individuals get the fitness benefits due to belonging to the group, individual selection overrides the group selection whenever individual and group selection pull evolution in different directions. If only behaviour that maximizes the individual fitness is selected, there is no need for group selection in models and explanations. Much of sociobiology aimed at showing how social evolution can be understood while staying in the individualist framework. An individual's fitness, at least in an inclusive sense (which then became the "gene's point of view" of evolution; see Dawkins 1976; Sterelny & Kitcher 1988), became the universal viewpoint for evolutionary explanations. Altruism became a central theoretical problem for sociobiology, but in most cases it was assumed that altruistic behaviour could be explained through *kin selection* (Hamilton 1964a&b) as maximizing one's own genes' fitness through relatives that share the genes, or as *reciprocal* helping (Trivers 1971). In both cases, altruism was only "apparent" from the evolutionary point of view.

The renewed idea of group selection has since been re-introduced by a handful of biologists and philosophers (see for example Wilson 1975 & 1989; Sober 1984a; Heisler & Damuth 1987; Damuth & Heisler 1988; Goodnight *et al* 1992; Wilson & Sober 1994; Goodnight & Stevens 1997; Sober & Wilson 1998; Okasha 2006; Goodnight 2012). Some evolutionary human scientists (for example, Boyd & Richerson 1985 & 2005; Gintis 2000b; Henrich & Henrich 2007; Bowles & Gintis 2011) have taken this approach seriously, but it is still not a widely, let alone a universally, accepted idea. There are several good reasons for this: the apparent weakness of group-level selection in competing with individualist counter-selection, the possibility of alternative (seemingly) individualist models when group selection seems to take place, and difficulties in finding unequivocal empirical evidence for it (for example, Kerr & Godfrey-Smith 2002; West et al 2008; Gardner 2013). I will

argue that some of the discussion on group selection is, after all the philosophical scrutiny over it, still somewhat ambiguous, and argue that group selection would be, in some of its senses, credible in the human context even if it was rare or non-existent elsewhere. But there are also human-specific concerns about the applicability of evolutionary explanations to human social behaviour in the first place, given the plasticity of the human mind, the power of culture to guide behaviour, and the novel complexity of human societies. I will not argue for the relevance of evolutionary approaches in this dissertation, but I will next discuss this topic and provide arguments for why such approaches may at least be worth trying.

1.2.1. Evolutionary Explanations in Human Context

Humans are biological beings whose basic characteristics are products of an evolutionary process. These include human behaviour and culture – they are biological phenomena. This claim does not imply anything specific as such, for example, about the freedom of the will; fixedness or plasticity of behaviour, or its development; the range of possible cultural variation in behaviour; the applicability of evolutionary models to contemporary human behaviour; or the relevance of primatology or any other field of biology to human social or behavioural sciences. All these issues depend on what *kind* of biological beings humans are.[8] The evolutionary origin of human social behaviour would be a theoretically important issue even if it did not have any consequences for any other scientific pursuits that build our understanding

[8] Neither is this an issue about whether humans are "blank slates" or not (Pinker 2002). Human mind has gene-related evolved developmental tendencies that guide, skew, and constrain the psychological make-up we can end up having as the end result of psychological development, but this general fact alone leaves open a wide range of possibilities from innate, massively modular mind to high levels of developmental plasticity. I will return to these issues later.

of ourselves. The knowledge of the origins and the naturalized version of "design" is an essential part of what can be called "naturalized philosophical anthropology", a science-based (but not necessarily reductionist) view of humanity, and of humans as "fundamentally social animals", as it were. Whether or not the evolutionary approach has more concrete consequences for human sciences depends on what the actual evolution has produced. The evolutionary history of the human species is what determines how much our knowledge of this very history helps us to understand the social behaviour of contemporary human beings.

There is a range of theoretical possibilities on this issue. On one end of the spectrum, evolutionary history does not have any consequences for any other epistemic pursuits about human beings. On the other end, the evolutionary perspective is a central organizing principle and a set of heuristics for generating fruitful research questions and hypotheses within other human sciences. Which theoretical possibility is the actual case, is, however, something we cannot know *before* the very research that evolutionary approaches are supposed to help. This is because of two things. First, given the epistemic constraints we have in tracing our evolutionary history, figuring out the past involves finding out what has evolved: it involves not only theoretical knowledge on evolutionary processes in general and evidence through fossils, but also some knowledge of the end-product, as well as comparative studies on the variation in contemporary humans and in our closest living relatives. Discovering our evolutionary history is dependent on our systematic knowledge of contemporary humans. Second, we cannot know how useful the heuristics based on the evolutionary "reasons" for our behavioural dispositions are in the process of building this systematic image of humans in contemporary settings, except by *comparing* the speculations based on evolution and the eventual empirical discoveries. We need to see where evolutionary speculations lead us, what we discover empirically, and how often the former helps with the latter.

In other words, whether evolutionary approaches are helpful for other fields of inquiry or not, is already a *substantial hypothesis* about what human beings are like and what their evolution was like, and this cannot be known prior to research. On one hand, we cannot draw the evolutionary picture without the systematic picture, and on the other hand, theoretical considerations alone, be they philosophical or evolutionary, cannot resolve the issue of how much help the evolutionary history provides us for drawing the systematic picture. Consequently, if we suspect evolutionary considerations to be useful in generating knowledge on human behaviour, we need to build both pictures – evolutionary and systematic – in close interaction. Since the usefulness of evolutionary approaches depends on what kind of beings we are, which is discovered through empirical research, and since we want to know whether the evolutionary approach is helpful in this empirical project or not, the only way to approach the question is through testing evolutionary ideas in empirical research. Hypotheses that are inspired by evolutionary considerations and the speculative history of our species may or may not turn out to be an important and fruitful part of this empirical project, but testing these ideas is the only way to know this.[9] The ideas to be tested include both the known details of the *actual human evolutionary history* and the knowledge about the *kinds of causes* that guide evolutionary processes in general for hypothesizing what psychological and social phenomena (and the proximate mechanisms producing those phenomena) could, could not, should, and should not exist. If this strategy for generating hypotheses turns to be successful, we have implicitly produced evidence for evolutionary history being important to understanding our contemporary selves.

The above is, in part, why I do not evaluate the evolutionary human behavioural sciences as such but rather concentrate on some

[9] This is a variant of the Lakatosian idea of how the core assumptions of research programmes get tested: not directly, nor by theoretical arguments, but through their usefulness in research (Lakatos 1970 & 1978).

methodological issues of what this science should be like if it is going to be successful. Philosophers have contributed their criticism volubly (and deservedly) in general, and I have little new to add. Instead, I will try to make a positive contribution to the theoretical basis. Even if the evolutionary social science turns out to be impossible, the same issues will still turn up in the study of the "mere" evolutionary history of human sociality, and evolutionary history and its guiding principles are interesting enough issues even as stand-alone questions. The topics I am exploring have consequences for this stand-alone project, at the least. They will be, when combined with direct empirical knowledge of human behaviour and psychology, a part of our understanding of what human sociality and culture are in the first place.

Most scientific approaches to human behaviour are interested in proximate causes. The developmental perspective, however, is at least an important heuristic regarding what behavioural dispositions there can be (they need to be possible to develop, after all) and what external factors the individual is sensitive to. For example, all theories that presuppose culture to have something to do with differences in human behavioural dispositions also presuppose that some cultural entities (whatever they are) play a causal role in individual development. The evolutionary perspective, in turn, focuses on the history of the behavioural disposition on a population level. An evolutionary history of a trait is essentially a descriptive history, which includes, to some extent, explanatory factors that tell us *why* certain behavioural dispositions exist in the population rather than some other dispositions. *At minimum*, then, evolutionary knowledge increases our self-understanding by telling us how the phenomena under investigation came to be, historically. This in turn may include understanding the constraints and scope of the behaviour. *At maximum*, the evolutionary history of a behavioural trait can highlight the function of behaviour in respect to its environment in a way that links certain environmental factors with the behaviour. If this is the case, the evolutionary perspective tells us why the behavioural disposition under study is coupled with certain environmental stimuli (including stimuli in the social

environment) and why certain environmental factors (including factors in social and cultural environment) have the influence they have in the developmental process. This knowledge can also help us discover the range of environmental variation in which the behaviour can be expected to emerge, by turning our attention to what is relevant in the environment (see Narvaez *et al* 2012 & 2014, for example).[10] If some position at the maximal end of these possibilities proved to be true – if we could rely on the connections between environment, behaviour, and the evolutionary function to hold – then we could turn evolutionary speculations into useful heuristics for hypothesizing about proximate mechanisms and behavioural phenomena.

To use an example of human social intelligence to illustrate these ideas, an example of a "minimal" project would be something like some of the more general explanatory work on what is called "Machiavellian intelligence"[11] (Byrne & Whiten 1988; Whiten & Byrne 1997;

---

[10] There is a net of proximate causal factors (both internal and external to the individual) that bring about the external behaviour in the context of the actual behaviour. Another net of causal factors (both internal and external to the individual) guides the development of these behavioural dispositions. Yet another set of causal factors explains the evolutionary history of those dispositions and interactions between an individual and its environment in both the proximate causal context of behaviour and its development. Any causal factor of the same descriptive type (e.g. a specific feature of environment) may be an explanatory factor in all dimensions. For example, in the human context, a type of interaction with other human beings (including cultural transmission) may be a part of the causal environment of the actual behaviour, of the development of the behavioural dispositions, and of the selective environment during the evolution of behaviour to the extent that it is similar to the ancestral social environment. The functionality of a trait may depend on these factors to be the same. If the environment changes from the evolutionary context, the functionality, the development, or both may change as well.

[11]The basic idea of "Machiavellian intelligence" is that the roots of human social intelligence (and primate intelligence in more general – the term was introduced by Frans de Waal (1982) for a chimpanzee social intelligence) lie in

Byrne 1997; see also Corballis & Lea 1999). A "maximal" project on the same issue would be to use the knowledge of or speculations on the context of the evolutionary origins, the models related to that context, and the plausible predictions they generate, as a basis for directly testable hypotheses on psychological capacities and biases, such as the evolution-based work on the "Wason Selection Task" (Wason 1966) and its connection to social adaptativeness in cheater detection (Cosmides 1987; Cosmides & Tooby 2005b).[12] Both approaches may be valuable at times – it is a trait-by-trait issue, not an issue with a general answer.

Even for the "mere" understanding of ourselves (or, the minimal contribution), the evolutionary perspective is not only an extra dimension to our self-understanding, but a fundamental piece of knowledge to satisfy the intellectual curiosity that often drives scientific and philosophical pursuits. It makes things (that would stay either messy or mysterious without the evolutionary perspective) intelligible. As Theodosius Dobzhansky (1973) famously put it, "nothing in biology makes sense except in the light of evolution." This quote should not be taken as a point about our inability to understand how the proximate mechanisms work in biology without the evolutionary perspective – we are perfectly capable of doing so – but the evolutionary perspective enables us to make sense of *why* the things are the way they are. Where did the apparently purposeful, often holistic, design come from? Why are there dysfunctionalities and faults in this design, from an intelligent designer point of view, that tend to be similar in kind to

---

the evolutionary "arms race" of social cognitive capacities in the environment of social competition, politics, coalition building, and manipulation.

[12]The idea here is that the capacity to formally infer along the lines of a material implication has been selected for detecting cheaters of social norms or contracts and therefore is activated in contexts involving social norms, making the inference practically automatic in these contexts, whereas in non-social contexts people make mistakes and find the formally similar inference tasks much more difficult. This hypothesis has a number of testable consequences. For criticism of the idea, see Davies et al 1995; Sperber & Girotto 2003.

other such faults – often as if some parts were taken from another design? Why are there clusters of characteristics across groups of species without any apparent reason for these characteristics to cluster? The same applies to human social behaviour to the extent that we can understand it from an evolutionary point of view. When we can cast an evolutionary light on it, we start making more sense of it. For example, suppose we discovered the evolutionary origins of morality. That is, we could explain why we are interested in the categories of right and wrong in the first place, and why we feel that we need to act in accordance with what we perceive to be right. This would deepen our understanding of ourselves even if it had no consequences for our moral practices or ethical theories – or even for our scientific understanding of moral psychology. If we did not have even plausible speculations about its origins, it would be mysterious and possibly, although not necessarily, a challenge to a naturalistic world view as well.[13]

1.2.2. Evolution as an Integrative Perspective

There is, however, a more important way in which evolutionary perspective strengthens our understanding: *scientific integration.*[14] Evolutionary perspective provides a framework for understanding how

[13] Even a plausible speculation may do some important intellectual work on this even if we had no way to confirm our speculations. William Dray's idea of *how possible* explanations in history proper was exactly that even though we cannot always provide verifiable explanations for historical events, we can make them *intelligible* by producing a narrative that makes them fit with what we know (Dray 1963; see also Persson 2012). The event or phenomenon under interest ceases to be a mystery or an anomaly, even if we do not know if the provisional explanation we give it is correct or not. I will briefly return to how possible explanations in chapter 3.

[14] This is also discussed in Dobzhansky's 1973 paper, although from a pedagogical point of view.

various properties of an organism and its behaviour are functionally related to each other. Consequently, this perspective makes intelligible how the different *fields* of biology are related, through giving a perspective on how their *explananda* are related. This integrative understanding comes from the idea that biological organisms are functional wholes that have parts and behaviours that relate to each other *as if* the organisms had been designed to function in certain ways in certain environments. This *design stance* (Dennett 1989 & 1995) toward biological organisms is made possible, justified, detailed, and constrained by the evolutionary histories of biological entities.[15] In the case of human behaviour, this turns into *interdisciplinary integration*. This has been an important piece of rhetoric behind much of evolutionary human science. Evolutionary psychologists of the nativist school in particular (see Tooby & Cosmides 1992; Cosmides & Tooby 2005a; Pinker 2002) have been optimistic on this point, expecting the evolutionary theory to become the foundational theory of all human science: concepts, theories, models, and methods from evolutionary biology would be a reductive basis for all human research. This kind of *reductive unification* would be the theoretically maximal contribution of evolutionary biology to human sciences. The idea has many problems, some of which I will discuss later. However, there is a less ambitious and more realistic variant of evolutionary unification that gives evolutionary approach a special role.

Even if the proximate-level approaches to human behaviour remain independent of evolutionary theory (and presumably of each other as well) both in their content and in how to characterize the proximate mechanisms, the evolutionary approach can nevertheless provide an integrative perspective on how the phenomena are related: it can be a tool for a *pluralistic integration* instead of *reductive*

[15]However, the design stance, if adopted, and its usefulness need not coincide with the actual evolutionary history or be adaptationist throughout. I will discuss this in chapter 3, but for the overall discussion at hand, this issue can be postponed.

integration. *Scientific pluralism* is the idea that even if the world is ontologically unified,[16] our epistemic practices are constrained to fragmented partial perspectives (either *de facto* or necessarily), and therefore our best theories and the knowledge produced by our best scientific practices could not be neatly integrated even if they were all true. Acknowledging this should also guide our epistemic practices. Pluralism has been developed by Helen Longino in particular (1990, 2002 & 2013). The idea of *integrative pluralism* is argued by Sandra Mitchell (2002 & 2003). According to this view, since reality is ontologically unified, even if the knowledge produced by different epistemic projects could not be simply added up, they refer to fragments of the same phenomena and we should be able to integrate these fragments of knowledge in practice. The locus Mitchell gives for this integration is explanation. Although she would probably disagree with the special role for the evolutionary approach, I consider my stance on explanation in general, and biological explanation of behaviour in particular, to be that of integrative pluralism.

Both Longino and Mitchell pay attention to explanatory interests and practices in their discussion. I agree that explanation is the key to understanding the connections and disconnections between the "epistemic units of science". However, I would argue that the relevant epistemic units for integration must be wider than individual explanatory models (Mitchell's focus) yet smaller than disciplines: the prime epistemic units of science are *local epistemic projects* that use models, theories, and other epistemic tools in an integrated manner. Individual explanatory models cannot be isolated from their explanatory practices and wider contexts of use, including substantial presuppositions, when compared. Disciplines, on the other hand, are far too broad and heterogeneous to function as epistemic units, which is reflected by Mitchell's examples from biology. If the evolutionary approach will help in integration in any practical way, it will be by

[16] John Dupré (1993) would be among the very few philosophers disputing even this.

providing a theoretical perspective to more local integration rather t han a general conceptual or theoretical framework that spans disciplines. The integration by evolution needs to be done separately in every case, and in the substance, not by re-conceptualizing behaviour and psychology in evolutionary terms, for example. However, I will argue that this can sometimes be done, and I will present a framework for doing so, based on the New Mechanistic Philosophy.

To sum up the above with an example, consider an evolutionary account of the foundations of culture (for example, Boyd & Richerson 1985 & 2005; Sperber 1993) and the evolutionary reasons for plasticity in behaviour and its development (West-Eberhard 2003). They transform the juxtaposition of the cultural and biological into an integrated image. This is enough to increase understanding on the philosophical level, and therefore evolutionary considerations should not be overlooked in social theory or philosophy. Nevertheless, this alone does not imply much in terms of empirical inquiry. Something more substantial is needed. One direction is to use the evolutionary perspective as a framework for thinking about the division of labour between various explanatory projects within human sciences. This could generate new research questions and hypotheses in interdisciplinary contexts even among "traditional" human sciences. Even if evolutionary considerations of this kind have, in the end, little to offer the actual scientific practices and substantial theories about the nature of phenomena under investigation, evolutionary understanding of the nature of culture and sociality would still have theoretical significance and could be used as an organizing perspective on human sciences.

1.2.3. Evolving in Groups

One of the most striking changes in human species during the period that most shaped human mind, from the early *homo* species to the emergence of anatomically modern humans with advanced tool cultures, is the increased intensity of sociality and the various interconnected, co-evolving phenomena related to it. These include collective

hunting and gathering; sharing food (as well as other goods and information) in growing frequency; increasing male participation in raising children (which also promoted monogamy, egalitarianism, and skill-based authority instead of individual-based hierarchy); more sophisticated tools and weapons requiring cumulative culture, which in turn requires (or is at least reinforced by) social learning; the increasing importance of social learning in general, which requires prolonged childhood, and a bigger brain that requires being born immaturely, both implying personal vulnerability and, probably, collective upbringing of children, including support from individuals of post-reproductive age, resulting in exceptionally long lifespan; folk psychology and language (as cognitive capacities and cultural media) to enable and intensify the above phenomena; social norms and sanctioning for cheaters; long-distance trading; pro-sociality and proto-morality, eventually full-blown morality; and many others.[17] Humans were essentially evolving in groups. They were adapting to their environments as *social collectives* much larger than extended families, with *cultural differences* between them, and these groups had relationships with other groups that transcended mere relations between individual members of the respective groups. Instead of all this sociality and culture being an argument against evolutionary approaches to human social behaviour, I take this to be an argument to approach the groups as evolving units in human evolution. The design that the social behaviour is a part of belongs to the *evolutionary design of the group*, and the adaptationist heuristics to human behaviour as well as to

[17] I will not provide a review of the vast literature on these issues here, but data and discussion on various aspects of growing sociality and the interaction between its different aspects can be found, for example, in Bar-Yosef 2002; Boehm 1999; Boyd & Silk 2003; Byrne 1997; Cela-Concode & Ayala 2007; Corballis & Lea 1999; Enfileld & Levinson 2006; Foley & Lahr 2003; Hatfield & Pittman 2013; Henshilwood & Marean 2003; Kaplan et al 2000; Levinson & Jaisson 2006; Lewin & Foley 2004; McBrearty & Brooks 2000; Mithen 1996; Tomasello 1999 & 2009; Roebroeks 2007; Sterelny 2012 & 2013; and Whiten & Byrne 1997.

psychology should take this seriously – as I will argue during the course of the dissertation.

The evolution of individual capacities and behavioural tendencies took place in a highly social context, partly for the needs of intensified group living. It is more than plausible that some of the behaviour has been selected for its social consequences that are not perceived by the agents and that are non-consciously triggered by social factors. This means, I will argue, that individualistic conceptualization of this particular kind of behaviour is not adequate – one has to go both beneath and above the individual level. The evolutionary "design" exists partly on the group level. It is likewise plausible (and there is backing empirical evidence) that a great part of what guides our behaviour is based on attitudes, norms, behavioural scripts, or some other behaviour-guiding psychological entities that are acquired from others during individual development through a process that cannot be described as reflective learning, and both integrates the behaviour of individuals in the group and separates it from other groups. An additional question is: can all this be understood as a product of individualist selection processes? I will argue that although there is no conceptual or theoretical necessity, both of these other forms of holism are connected with the group-level perspective in the evolutionary explanation of behaviour and arguments for group selection in explaining the evolution of human social behaviour. Furthermore, I will argue that these two aspects – group as an interactor (the proximate dimension) and group as a replicator (the developmental dimension) – imply different kinds of group selection processes, although mutually enforcing ones, that should be treated separately in the debates over group selection.

1.3. Individualism and Holism in the Evolutionary Social Science

All behaviour, including social behaviour, depends on psychological capacities and tendencies, but not exclusively. Even if mind's structure was modular[18] and each module could be identified with a specific adaptive function, as some evolutionary psychologists think (for example, Cosmides & Tooby 2000 & 2005a; Buss 2003), the behaviour that these modules generate in contemporary environments might be very different from the behaviour they produced in the environments of their evolutionary origin. Furthermore, independently of the previous point, even the same behaviour might have different functions in the overall behaviour of an individual in the current social complex than it had in the adaptive context. Consequently, the relevance of evolutionary considerations seems to be limited to what can be said of individuals. This is not necessary, but even if it were true, it would not imply that evolutionary considerations should be individualistic in any other way. That is a further assumption. Furthermore, there are two versions of holism, *interactive* and *collective*, in each of the dimensions.

1.3.1. Individualism, Interactionism, and Collectivism

The central idea of my approach is the following. The question about explanatory individualism is the question about the *explanatory relevance* of the properties on different levels of biological organization.

[18] That is, the structure of cognition is such that there are functionally independent cognitive capacities that take the input information from a limited number of other modules, process it independently and mechanically, and then put forward output information to a limited set of other modules, some of the modules taking information from perceptual organs and some having bodily behaviour as the outcome of processing. (For the idea of modularity, see Fodor 1983; Sperber 1996; Carruthers 2006; Anderson 2007; for criticism, see Karmiloff-Smith 1992 & 2006; Woodward & Cowie 2004; Buller 2005.)

This is a local question and may have different answers regarding different explanatory interests even about the same biological phenomenon. Furthermore, there are three dimensions along which the evolutionary explanation may or may not be individualistic or holistic: proximate, developmental, and evolutionary. The explanation is individualist in the proximate dimension, if it requires considerations about the individual-level behavioural traits only (and the environmental context, including the behaviour of others, is treated as the selection environment only) and about the proximate mechanisms underlying this behaviour on the individual level, and those individual traits are the target of the evolutionary explanation. The explanation is holistic if the assessment of evolutionary functionality requires supra-individual attributes (for example, social mechanisms and structures, group properties, or something like that) to be the object of selection. That is, the social-level properties are not just a *consequence* of the individual characteristics that are being selected, but there is a selection between *forms of interaction* or *interactive phenotypes* (Moore *et al* 1997; Wolf *et al* 1999) and their consequences, or between group traits.

An evolutionary explanation is individualistic in the developmental dimension if all the components of the development that are relevant to explaining how these behavioural capacities came about are insensitive to the relevant variation in social and cultural surroundings. This does not require the development be *causally independent* of social factors but that the same behavioural tendencies emerge in any social and cultural environment. In other words, an *innate* psychological trait[19] is an individualist trait, and individually learned behavioural tendencies are individual traits. But if the

[19] I will return the concept of innateness in detail later. An archetypical individualistic evolutionary approach to human behaviour in this dimension is nativistic evolutionary psychology, and I will argue that there is an adequate notion of innateness that captures what nativistic evolutionary psychologists consider to be their object of research. Whether this is an adequate research paradigm or not, is another issue. My interest here is only in explicating what this approach is about, in a charitable reading.

development is dependent on specific social or cultural variables that may vary from group to group, the development – and therefore the reproduction of the behaviour – has a holistic component. And, similarly, in the evolutionary dimension, the individualism-holism issue is defined by whether the explanation by fitness consequences requires only differences in properties on the individual level or also group-level differences. I consider all these issues to be methodological, pragmatic issues, not ontological. The relevance of the group level in any of these dimensions requires a holistic approach of some sort in the evolutionary explanations. I will discuss the levels of selection issue in the last chapter of this dissertation. I will argue that the proximate and developmental dimensions have direct consequences for the levels of selection issue in the evolutionary dimension and that this is already implicitly a part of the levels-of-selection discussion. However, not distinguishing between the three different dimensions leads confusions. Moreover, there are different *degrees* of holism in each of the dimensions. At the extreme end, the traits and processes of interest are on group level exclusively. I call this **collectivism**. However, the relevant traits and processes can be neither individual or collective-level – they may be "in between" the levels, emerging in the interaction of individuals in the group structure without being reducible to individuals only. I call explanations referring to such properties and processes **interactionist**.

Generally, I will call the assumptions that individual perspective is what we should be exclusively interested in, when it comes to evolutionary explanations, **explanatory individualism**. To give a preliminary characterization of pure individualism, it is a view in which social behavioural traits

Ind.1 can be adequately described in terms of individualistic goals and/or behavioural dispositions and mechanisms for evolutionary explanatory purposes,

Ind.2 are innate, and

Ind.3 have functions that they have been selected for that are exclusively for the benefit of the individual's (inclusive) fitness in the social context of the selective environment.

The *polar opposite* of individualism is **collectivist explanatory holism**. A preliminary characterization for the three collectivist dimensions is this:

Col.1 Behaviour and the underlying mechanisms are fully socially contextualized in their function. That is, the proper function of the behavioural response to a social stimulus, for explanatory purposes, is on the collective level, not on the level of goal-oriented individual agents. The behavioural traits are traits of a collective (a group).

Col.2 The behaviour and its underlying (psychological) mechanisms develop in interaction with the socio-cultural environment in a way that makes them highly culture-dependent: we are constituted by the culture we grow in. In order to understand social behaviour, we should understand culture and social structures, not psychology.

Col.3 All this has been evolved on the group level and for directed, functional plasticity: there has been group selection between social behavioural traits that has de-coupled the social level function and individual goals and has directed the learning biases and heuristics in social learning in a way that makes cultural groups cohesive adaptive entities, individuals in their own right.

The last claim is not derivative of the first two. Behaviour could have functional social-level consequences not reducible to goal-oriented actions of the agents without this being the locus of selection; even if selected, the selection could be individual-based; and the behavioural plasticity of social learning and culture could have a general evolutionary function (that is understandable through individual selection) instead of group-related functionality. If human groups were *superorganisms*, not just collections of individuals, the full-blown collectivistic

holism would be true in general in all dimensions. This is not the case, but there might be some traits that are collectivist in this sense, in one or more dimensions.[20]

I call the middle range option **interactionist explanatory holism**. It can be characterized (preliminarily) along the three dimensions as follows:

Int.1 Social behaviour has both causes and consequences that are not a part of agent's individualistically definable goal-oriented psychology but have a (selected) social function that is not reducible to the agent's own goals. *The social behavioural traits are interaction types between individuals.*

Int.2 Social behaviour is acquired from others, but it is not uniform across the collective.

Int.3 The selection between behavioural traits is between groups of interacting individuals within the collective – neither between individuals nor between collectives.

I call this *interactionist* explanatory holism, since the focus is not on groups as cohesive collectives, but on the kind of interaction between individuals in a social context that is not reducible to individual behavioural dispositions alone. The views I mostly explore in this dissertation are interactionist, and the point is to contrast them it with methodological individualism. If the term "holism" is used without qualification, I am referring to the interactionist form of anti-individualism. I will discuss the more precise contrast between collectivism and interactionism and their consequences more concretely during the discussion of each dimension. The intuitive idea, however, is this. The whole group of individuals sharing space and interacting with each other is the group as a **collective**. Humans were organized for most of our species history into these concrete, distinct groups. If this is the level at which the relevant explanatory factors take place, we are

[20] But see Pagel 2012 for a view that comes very close to a collectivist, superorganism view of human cultural groups and human evolution.

giving a holistic explanation. Some individuals interact with each other more than others within the collective. These connections or the networks they constitute may be the relevant explanatory factors instead, in which case I call the explanation an interactionist explanation. These individuals form an **interaction group** based on the specific forms of interaction that are relevant to the explanatory purposes. Interaction groups are constituted through specific, temporary interactions, while collectives are the static collections of individuals who interact with each other in multiple ways. If an interactionist trait becomes collective-wide, it becomes a trait of the collective. There are significant differences between the two ideas of group. For example, there is competition between interactionist traits within the collective and the interaction groups overlap, but the competition between collective traits is between collectives only and the collectives are distinct. I will discuss the different dimensions separately, but the first two (proximate and developmental) constitute *conditional parts* of an argument for group selection; given other factors, both proximate and developmental holism about a trait can be a reason for using a group selection model for its explanation.

1.3.2. An Overview of the Dissertation

Asking a question the right way and using the right kind of tools in answering it play a big part in giving an answer to a philosophical question. The structure of this dissertation reflects this. The first half or so will be preliminary work for the main topics, but this work sets the stage for them and articulates crucial background assumptions and premises, as well as develops some parts of the arguments already. I will begin the dissertation by explicating what I mean by the evolutionary explanation of behaviour. My perspective is the *contrastive-counterfactual theory* as the *normative theory of explanation* and *mechanistic philosophy* as a way to understand how things are connected in explanatorily relevant ways. My interpretation for what counts as a mechanism is relatively liberal. Mechanisms do not need to be *spatially*

connected and restricted structures, for example – the important thing is that there are *causal interactions that form functional structures*. I will build my arguments directly using these ideas. I will, therefore, devote much of the first substantial chapter of the dissertation to explicate how I understand these approaches, especially in connection with my main subject matter. I will not provide the view a systematic defence, since this is only background theory and both views (contrastive-counterfactual theory and mechanistic thinking) are mainstream, even dominant views. But I will spell out some insights of these approaches that inform my further discussion and may be controversial.

Under the contrastive view of explanation, the issue of individualism and holism itself is about which levels of biological organization have *explanatory relevance*. This is not an ontological, but a methodological issue – how to break the causal processes or the mechanistic structures into explanatory parts in an adequate manner. There are pragmatic criteria for considering some subset of the causal factors relevant for explanation, in a relevant balance of accuracy and robustness in description, and the main criterion is whether there is variation within the factor in the contexts relevant to the explanation, such that it is coupled with variation in the explained behaviour. The rest of the actual causes can be considered as fixed causal background or noise. This applies to all explanatory dimensions. The questions to be asked about individualism and holism are, therefore, about what is the relevant level of description to pick causal factors for behaviour: individual or social. From the evolutionary point of view, the explanatory variation within social environment is interesting to the extent that it is relevant to evolutionary explanation. This has three components: what is evolving (that is, if the property being selected for is a property of an individual or of the group); what is the underlying replication process (that is, whether the processes transfer the property through an individual or social-level process), and what unit of biological organization we should attach the fitness consequences to.

I will discuss evolutionary explanations of behaviour and the relationship between evolutionary explanations and other biological

explanations in the third chapter. I will start by discussing the adaptationism debate and arguing that there are three different sensible interpretations of adaptationist explanations: *historical adaptation explanations*, *ahistorical adaptive function explanations* and *current use analysis*. Explanations in the evolutionary human sciences can be any of them, and different ways of doing evolutionary social science use different explanations, as I will show in chapter 4. These are all forms of what I call *evolutionary functionalism*. The logic of explanation is the same in all of them, and most of the other issues I discuss apply to all of them. They face different constraints, however, and the contents of the explanatory questions and the explanations are distinct. I will also outline a model of evolutionary explanation of behaviour that I call the *mechanistic view of evolutionary explanatory hierarchy*. This will be the perspective from which I will connect the proximate, developmental, and evolutionary dimensions of individualism versus holism issue. The interaction between these explanatory dimensions has consequences for how holistic connections in proximate or developmental dimensions affect the question of levels of selection. I will then briefly discuss evolutionary social sciences (*sociobiology*, *evolutionary psychology* and *evolutionary anthropology*) in chapter 4 to provide background and to contextualize the main topics within the evolutionary human sciences as they are practiced.

After this I will move to my main issues, starting with the proximate level. I will distinguish between behavioural and psychological traits as different *explananda* in chapter 5 and examine their relationship. I will later argue that some behavioural traits should be understood as interactive traits. I will spend much of the chapter 5 discussing *folk psychology*. I need a way to characterize behavioural traits independently of the psychological architecture underlying it, and for this I need an account that specifies what the behaviour is *about* without an intentionalist baggage. Our folk-psychological practices approach behaviour as action that is guided by the individual's goals and beliefs about means to achieve those goals. This builds a direct link between behaviour and psychology: the content of behaviour is

seen in the *reasons for action*. I will argue that this is an inadequate way to approach either psychology or behaviour for the current explanatory purposes and I propose an evolutionary functionalist analysis for the individuation instead. I will not present an eliminativist argument, however – I will argue that folk psychology *conflates* proper psychological and (what I call) *agentive* descriptions. The reference of folk-psychological concepts is complex, not non-existent.

I will discuss individualism and holism in the proximate dimension in chapter 6. However, I will mostly discuss various concepts of altruism in this chapter, both as a continuation of the previous chapter and to explicate the idea of *interactive behavioural traits* by taking reciprocal altruism as an example. I will distinguish between evolutionary, psychological, behavioural, and agentive notions of altruism, partly based on the discussion of the previous chapter. I suggest that the vernacular notion of altruism in human context conflates psychological and agentive notions, but that agentive altruism is the central concept for practical purposes. Furthermore, agentive and behavioural concepts may be confused in describing helping behaviour. I will also discuss the concept of evolutionary altruism and relate it to behavioural altruism in biological contexts. I will then analyse reciprocal altruism and argue that all four concepts of altruism are relevant in understanding this in the human context, and that there are two different traits to be explained. What is selected is not *only* the disposition to engage in certain forms of interaction, but the *forms of interaction* that get selected against other forms of interaction, firstly, and the disposition to engage in such interactions, secondly. Furthermore, the underlying psychology and the forms of behaviour are also different in the sense that the same psychology can instantiate different behavioural patterns depending on the social environment. If it is a form of interaction that brings the fitness benefits, this is what gets selected, and the psychological capacities to participate in such interactions are selected based on this.

I will then move to the developmental dimension. This is an important dimension for understanding evolution in general because

what gets selected is not the disposition for the behaviour, but the developmental process producing the disposition. Much of human psychology and behaviour is learned, flexible, creatively improvised in the situation, and so on. Much of this cannot be an object of evolutionary explanation at all. There must be some processes of inheritance that 1) make the reproduction of the trait reliable enough to be an *object of selection*, and 2) are difference-makers between the *units of selection*. I will discuss the relevance of development for evolution and the *Extended Synthesis*, as well as the developmental individualism and holism issue from this perspective in the chapter 7. I will also briefly discuss the role of culture in all this as a medium of inheritance that facilitates holism. However, most of my discussion on development will be about the concept of *innateness*, which I will discuss as the form of evolutionary individualism in the developmental dimension. I will reply to the critical discussion on the concept of innateness by defining and defending a *contrastive invariance account* of innateness. I will also argue that a charitable interpretation of nativistic evolutionary psychology (or, alternatively, a normative recommendation for how to understand nativism) is to understand it as a *methodological choice* that constrains the applicability domain of such evolutionary psychological research. The aim of the dissertation is not to argue for holism across the border – I will argue for a neglected form of holism in the proximate dimension, but I want to retain individualism, too, as a sensible approach in the developmental dimension.

However, evolutionary psychological explanations do not need to be developmentally individualistic either. Since the context of (at least some crucial steps of) the evolution of human mind was already social and cultural, some of the species-typical individual development of mind may require certain socio-cultural elements in the developmental environment. If there is directed, functional plasticity that is guided by these elements, one could have an evolutionary psychological explanation that is not nativist (developmentally individualist). An example of this might be the psychological capacity for language.

The last substantial chapter of the dissertation addresses evolutionary holism in the purely evolutionary dimension – that is, group selection. I will also bring the other two dimensions into the discussion. Talking about groups can be misleading in this context: it is not clear what counts as a proper group. There are also different ways in which group selection can be understood. I will define evolutionary holism as a selection process in which the fitnesses of the individuals are tied to each other. This can take place in many ways: through a connection in the evolutionary dimension only (that is, the individuals affect each other's fitness because of the group structure, as in the case of evolutionary altruist groups outperforming evolutionary egoist groups), through a connection in the proximate dimension (the individuals participate in the social traits that have fitness consequences for the individuals, but only through participation), or through a connection in the developmental dimension (the reproduction of the trait is social). I will first discuss the issue of *multilevel selection* as a standalone issue, and especially the debate on kin selection as a form of group selection. I will argue that the selection logic in kin selection is that of group selection, although there are important additional elements to it. After this, I will show how holistic properties in the other dimensions promote group selection causally although they do not imply it. I will frame this discussion partly with the previous discussion on the various aspects of the levels of selection debate by Elisabeth Lloyd (1992, 2001 & 2017).

2. Explanation in Biology

An explanation of something (e.g. an event, some aspect of it, or a property of an entity) articulates why this something is, or why it is the way it is. I will begin this chapter with a brief discussion on what it means for an explanation to do so in general, and a causal explanation in particular. I do this partly to articulate the aims of the analysis, but the view adopted on this issue also has substantial consequences, as we will see when I take a brief detour to the debate on whether natural selection is a mechanism. I will later argue for a certain position on adaptationism that depends directly on how we understand the nature of explanation. Furthermore, this issue is relevant to how we understand psychological explanation, which will be a topic of some later chapters. After a preliminary discussion on the aims of theory of explanation, I will discuss the *New Mechanistic Thinking* in biology, which is a part of my general framework. After this, I will articulate a view of natural selection explanations that will inform the next chapter's argument for how to approach behaviour as an adaptation.

2.1. Causal Explanation

Philosophical theories of explanation aim to explicate what properly constitutes an explanation. At the same time, as an implication, they provide normative criteria for when something proposed as an explanation succeeds at being one, as well as criteria for comparing different explanations. The general idea is that explanation refers to a thing, the *explanans* (e.g. another event), which is in a proper relation to the *explanandum* (for example, there is a lawful connection between the types of things that *explanandum* and *explanans* are tokens of). Not all explanations are *causal explanations*, unless a causal theory of explanation proves to be a universally true theory of explanation, but for the

purpose of this dissertation, I discuss only causal explanation.[21] There are, however, two very different approaches to what the aims of a theory of explanation should be. I will call them *fundamentalist theories* and *pragmatic theories*. In the context of causal explanation, they also imply two quite different approaches to causal relations that can be characterized, likewise, as fundamentalist and pragmatic theories of causation. I will make some remarks on this now, before articulating the main ideas of the contrastive-counterfactual theory of explanation and the accompanying manipulation theory of causality.

2.1.1. The Aims of the Theory of Explanation

What I call the *fundamentalist theories* are those that aim to explicate how the *explanandum* fits to the fundamental structure of the world. According to some such views, the *nomological theories*, explanation involves general laws that describe or reflect the real regularities of the fundamental structure or processes in the world and from which the *explanandum* can be *inferred* (Hempel 1965; Salmon 1971; Friedman 1974; Kitcher 1989). According to others – call them *causalists* – explanations refer to ontologically fundamental causal powers and capacities, from which the *explanans* selects those that are relevant (Mackie 1974; Harré & Madden 1975; Salmon 1984; Cartwright 1994). According to these fundamentalist views, an individual object of explanation is fully explained when we know exactly how it fits to the way the world works in general. Discovering the fundamental change-

[21] A *causal theory of explanation* (e.g. Salmon 1971; Salmon 1984) makes the claim that a thing is explained when we know what caused it. This reduces all forms of explanation to causation, and by implication theories of this sort are theories of causation first and theories of explanation second. A *theory of causal explanation* (e.g. Ylikoski 2001; Woodward 2003), in contrast, is a theory that explicates a causal explanation without needing to be a theory of causation, to presuppose any specific ontology for causation, or making a generalization about all explanation being causal.

producing structure of the world (whether it includes only physical laws or also some emergent causal powers) is an important aim of science – whether we can do so or not. It is also an interesting philosophical question to ask what fitting an individual event into a fundamental structure like this would consist of. This cannot be a general theory of explanation, however, for the following reasons.

First, a theory that aims to explicate what takes place at the fundamental level of connectedness[22] between the things in the world is mostly unusable. It cannot be used to describe and normatively evaluate most of the actual practices of scientific explanation when, for example, we are interested in finding the right kind of systematic dependencies between two things of interest on a domain of research that is not at the level of these fundamental dependencies. In other

[22] There are multiple ways of understanding "levels", as discussed in the Introduction, and this applies here, too. One way to understand "more fundamental level" is to think about it in terms of ontological levels: some things function as the ontological constituent parts or "grounders" of the higher levels. There is a persistent intuition that more fundamental and more constitutive entities are also smaller: the entities of lower ontological levels are also the structural component parts of the systems of higher ontological levels. The idea of a mechanistic understanding of the ontology of levels (see Wimsatt 2007; Craver 2007; Levy 2013) sometimes rests on this intuition, but they are at least conceptually separate (Craver 2007 & 2015; Kuorikoski 2009). "Levels of explanation" can refer to either the hierarchical levels of constitution (like in the case of individuals and groups in the multilevel selection models) but also to the abstract levels of mechanism: the constituent parts of a mechanism need not be on the same ontological or hierarchical level with each other (Craver 2007). If this is adopted, what is "more fundamental" in mechanistic explanation is not the same as "more fundamental" in a structural or ontological analysis. Furthermore, levels may be levels of abstraction for analytical purposes even in explanation (like the functional and computational levels in cognitive science; see Marr 1982). The traditional way to think about fundamentality of levels in philosophy posits the fundamentality of "fundamental physics", which is usually thought to be fundamental in all the relevant senses. For the current discussion, I grant the existence of such a fundamental level, whatever this means. If there is no such thing, fundamentalist theories automatically fail.

words, it is not a suitable approach to explanation in special sciences, including biology. We need a philosophical theory for characterization and for criteria of evaluation here too. Second, things are connected in complicated systems in complicated ways, and looking for the fundamental connectedness (such as physical causal processes; for example, Salomon 1984) is not necessary, may even be misleading (see Woodward 2003), and does not reflect the explanatory practices in special sciences. When we deal with "imperfect" or "elliptical" explanations that do not give the whole picture, we need philosophical tools to compare different explanations, even if they are all true in their domain, to make sense of the whole picture and to relate the various explanations to each other. For example, we need tools to map the differences in what explanatory facts they point to and what the relation between these facts is. This calls for pragmatic theories of explanation – either to replace or, at least, to complement the fundamentalist theories.

The pragmatic approach to the theory of explanation takes the *practice* of explanation as its starting point for explication, clarification, and, subsequently, sophistication. It aims at a normative theory of a successful explanation without making a reference to a fundamental connectedness as a necessary criterion. This approach does not contradict the principal idea of fundamentalist theories that the real dependencies that the causal explanations trace are constituted by the fundamental structure of the world and how the change in it works – they are mostly neutral about this. It may be that some of the fundamental theories are compatible with some of the pragmatic theories, giving explications for different notions of explanation, instead of being rivals. One could, for example, consider pragmatic theories to be the *theories of "partial" explanations*, even if one thinks that a fundamental theory is needed to account for deeper questions of what ultimately makes things explanatory. The pragmatic approaches take the role of explanation, and therefore the role of the theory of explanation, to be more modest and practical than tracing fundamental structures exclusively. The notion of explanation in this dissertation is a pragmatic

one. I will not argue against fundamental theories, but they are irrelevant to my discussion.

The old school pragmatic theories were theories of the *pragmatics of explanation* (Gärdenfors 1980; van Fraassen 1980; Tuomela 1980; Achinstein 1983; Sintonen 1984). They aimed to explicate the explanatory questions and give criteria for practical evaluation of whether the given explanatory information was an answer to the explanatory question or not. Newer pragmatic theories, in contrast, aim at a normative theory of what *counts* as explanatory information and *for what*. The philosophical work in the newer theories is strongly based on the work done in the earlier theories, but the aims are different. Consequently, the new pragmatic theories include criteria for what kind of things in the world are explanatory – even if they are not theories of what make things ultimately explanatory on the fundamental level. There are two approaches to this. One is the *contrastive-counterfactual theory* (Woodward 2000 & 2003a; Woodward & Hitchcock 2003; Ylikoski 2001; Ylikoski & Kuorikoski 2010), which gives an account of how a partial explanation works, based on explanatory interest relative selection of *real dependencies*. The other is the *mechanistic theory* (Bechtel & Richardson 1993; Glennan 1996 & 2002b; Machamer, Darden & Craver 2000; Craver 2007), which aims to explicate the logic of explanation as a practice of revealing a type of dependency, a mechanistic connection that can be used in explanation.[23]

The contrastive-counterfactual theory and the mechanistic theory are not exclusive alternatives to each other but approaches concentrating

[23] Including mechanistic theories in this category might seem a controversial move: one could take the stance that mechanistic theory is a causalist theory of explanation, if one takes it to be a theory of causation, too. I will discuss this matter shortly. But the main reason for including it here is that the proponents of this theory are interested in the structural connections of mechanistic interactions in picking out the explanatory relations between things, instead of the causal powers of the entities that constitute these structures. Furthermore, the mechanistic ideas are used in this thesis as pragmatic, not ontological, ideas about explanation and causation.

on slightly different issues of causation and explanation, and they are often combined (for example, Craver 2007). This is a widely although not universally held position, and it is adopted in this dissertation. I take the contrastive theory to be the primary theory explicating the concept of explanation. The accompanying *manipulation account of causation* is the primary perspective on causal relations when it comes to the practices of causal explanation in science. The mechanistic theory is an important addition to account for relevant explanatory information and to speak about the nature of causal relations in biology (see Ylikoski 2001; Woodward 2002 & 2011). These approaches are not only mainstream in the philosophy of science in general, they are all but dominating views within the philosophy of biology, so I will not advance a systematic defence for them. It is, however, worth articulating what these theories say about explanation and some of the reasons why these are the adequate approaches to assume at this point, since these ideas inform the substantial discussion of the main topics.

2.1.2. The Contrastive-Counterfactual Theory of Explanation

The starting point of the contrastive-counterfactual theory of causal explanation[24] is that science traces real patterns of causal *dependencies*. The dependencies do not require a specific direct connection or productive relation between the cause and the effect. This idea does not contradict the idea that causal relations are *ultimately* constituted by the fundamental structure of the world and whatever connects events or things to each other spatially. The regularities that are supposedly described by the strict laws of future fundamental physics may provide the description on this level. For any object of study within special sciences, however, there is a multitude of causal connections in a complex network, and for any particular explanatory interest, only some subset of them is relevant. An explanation is an answer to an

[24] This brief review of the main points of the theory is based on Ylikoski 2001 & 2007; Woodward 2003a & 2004; and Woodward & Hitchcock 2003.

explanatory question that focuses on some subset and implicitly constrains what kind of answers are relevant. For example, if we are interested in knowing why a window got broken, we might sometimes be interested in the molecular structure of glass, sometimes in the intentions of the child who threw the ball at it. A complete causal description would include both (under some description) and a lot of other things, but that is not what explanations usually aim at. On the contrary, much (or even most) of the causal information is not even explanatory within the chosen explanatory framework. Real explanation is always aspectual. The contrastive-counterfactual theory analyses the logic of such a *partial explanation* and gives criteria for *explanatory relevance*. According to it, the proper logical form of an explanatory question is not just "why x happened", but "why x_1 happened instead of $x_2 \ldots x_n$", where $x_1 \ldots x_n$ are mutually exclusive alternatives. $x_2 \ldots x_n$ constitute the *contrast class* for the *explanandum* x_1 – therefore the explanatory question is *contrastive*. The contrast class may be implicit in the explanatory question, but it is analysable in the context, if the question is unambiguous.

For example, if we are interested in explaining a particular behaviour of an animal, say a cat attacking something small that moves, we are not interested in just *anything* that contributes causally to its behaviour, nor *everything* that is. We are interested in something that makes the animal behave in this particular way instead of some other particular ways. The contrasts are either real or theoretically possible alternatives (depending on the explanatory interest), and *mutually exclusive*. Knowing all causal dependencies relevant to the behaviour and contrasting it with all logically possible alternatives would maximize our explanatory understanding of the behaviour, but in practice, the explanatory interests are always narrower. The adequate alternatives could include things like the cat not paying any attention to the moving object at all or shying away from it. The explanatory factor we are after is the *difference maker* between these alternatives: it is the factor that leads into the *explanandum* instead of anything else in its contrast class. Different explanatory questions have different contrast

classes and call for different difference makers, even if they are all explanations for the same thing, such as the same behaviour. This is especially important in evolutionary explanations that explain a trait as being an adaptation for something. First, an adaptation explanation does not, at least in its proper use, simply say that a trait is *useful*, it says that it has been (1) *more* useful (2) for some particular purpose (3) than a set of alternative traits that either existed in the earlier history of the species or could have plausibly emerged instead of the actual trait that is being explained. Whatever the trait's use is, specified this way, it is a difference-maker. Secondly, an adaptation explanation needs not be a claim that this use was the *only* causally relevant factor in the trait's evolution – to give an adaptation explanation is to choose one property of the trait (its fitness-increasing use) special interest. Whether this is appropriate or not depends on the research question. There is a sharp distinction between using adaptive value as the explanatory focus (*explanatory adaptationism*) and assuming that the adaptive value is all we need to know to understand how the trait emerged (*empirical adaptationism*; see Godfrey-Smith 2001). I will return to this later.

The *explanans* must be contrastive too. Even if the *explanans* only mentions the difference maker, it presupposes mutually exclusive alternatives to it: the *explanans* is of a form "because y_1 happened instead of $y_2 \ldots y_n$". This is important for several reasons. First, for practical purposes, the description that picks the cause may be ambivalent regarding the details mentioned in the description. That is, the relevant causal factors of the *explanans* may not be specified by a mere reference to the *explanans*. Making the contrast class of the *explanans* explicit makes the relevant causal factors explicit as well. The property needs not be specified as such, only implied by the contrast class. However, there is a more substantial reason for doing this. In a complex system, several different changes in the causal history of the system could cause a similar change in the *explanandum* and different explanatory questions ask for difference-makers within different contrast classes. This must be made explicit when comparing explanations. Two

alternative causal explanations do not necessarily compete even if they point to two different factors that would make a difference in the same way. They may refer to different "locations" in the system that could be altered, and they may even all be necessary parts of the causal history of the system that brings about the *explanandum*, even with the same contrast class. This is important to note when comparing different explanations and explanatory approaches to human behaviour, for example: many of them may be complementary and true at the same time, despite superficial discrepancies, while all of them are partial and cannot be treated as more than that. I will return to this in later chapters. Adopting the contrastive theory helps to conceive this, and explicating the contrast class of *explanans* makes it explicit.

Take the explanation of the instinctive aggression of a cat for an example again. This behaviour as such is the complete target of our explanatory interests, and a full explanatory understanding of it would consist of knowing all the factors on which it depends, but any particular explanation points to a particular factor without which the behaviour would not take place. If we could intervene in the system such that the proposed explanatory factor is changed to one of its alternatives (without changing anything else), the behaviour would change,[25] but we are not interested in *any* such intervention. At least in principle, several different interventions in environmental factors, as well as in the animal's psychological (or neural) states, could have the same effects (similar changes in the behaviour), but these factors are not competing explanations for the typical behaviour, since they are searching for difference-makers within different contrast classes. In other words, they are mutually inclusive explanations and may

[25] It does not matter for the semantic analysis of the logic of explanation whether we could actually make these interventions – we are using the notion of *ideal intervention* in order to articulate the logic. The actual feasibility of the interventions matters only for whether we can discover the causal dependencies through actual experimental settings. The discussion for now is about the semantics of what is means to explain. I will return to the causality part of causal explanation shortly.

point to different factors in the network of causal connections that result in the behaviour. In addition to a direct intervention in the mechanisms triggering the behaviour, the cat's "complete" behavioural disposition could be affected through intervention in its *development*, which would not only be an intervention in a different factor, but in a different *kind* of factor. I will return to the various kinds of biological explanations later in this chapter. Again, the relevant change to the psychological (or neural) disposition could take place both internally and externally: by there being different genes or epigenetic differences in developmental pathways, and by the cat growing up in a different environment.[26] Another different dimension of causal explanation (with a different difference maker and contrast class) is called for if we are interested in the environmental factors in the evolutionary past that contributed to selecting for the behavioural disposition.

Many of the various difference-makers may make the relevant difference among the same alternative outcomes within the same contrast class of the *explandum* even if the explanations are wildly different. All true explanations pick up some factors that contribute to the behaviour being what it is, but the explanation is aimed at picking up an explanatory relation with the right kind of contrast class at both ends of the relation, and this is presupposed in the explanatory question even if it is not explicated. As pointed out above, no biological explanation asks for *all* causal factors at once (see Bechtel & Richardson 1993; Woodward 2010), and giving too much in an explanation makes it less informative and therefore a worse explanation for what is being asked.[27] Yet, at the same time, some explanations make

[26] The difference between contrast classes is determined by explanatory interests, which may depend on a practical reason for the explanation. For example, if one is interested in explaining species-typical regularities in cat behaviour, internal developmental causes may be more central, but if one is interested in affecting a cat's behaviour, external factors may be more interesting.
[27] One could argue, however, that the full understanding of any given event is constituted by all the causal factors that guide the event, i.e. an ideal explanatory text (Railton 1978 & 1981; see also Kitcher 1989). The Hempelian idea of

presuppositions that may contradict other explanations (Mitchell 1992 & 2002; see also Pigliucci & Müller 2010). Comparing explanations requires an analysis of a wider causal setting than an individual explanatory approach implies. This lies at the heart of the dissertation at hand – analysing the causal presuppositions of evolutionary explanations of behaviour that are not a part of the explanations themselves, such as presuppositions about the proximate and developmental causal processes.

2.1.3. Causes in Explanation

The contrastive-counterfactual theory of causal explanation is not only a *contrastive* theory of *explanation*, but also a *counterfactual* theory of *causal* explanation. According to it, *y* causally explains *x* if and only if a) y *was a part of the causal history of* x (as a matter of fact), and b) x *would not have taken place if* y *had not taken place* (contrary to the fact; Ylikoski 2001: 35; see also Woodward 2003a & 2004; Schaffer 2005).[28] In other words, a cause is something that the effect depends on. In this framework, causation refers to a *difference-making relation*, not a *process of production*. Moreover, it is not an *ontological* theory of causation, like fundamentalist causal theories of explanation, and it does not need to

---

giving all the relevant fundamental laws and conditions, or the mechanistic idea of describing the whole causal structure (this will be discussed shortly) may seem like natural ideas here, but they are both limited by only describing *positive factors*. Full understanding should also include knowledge about how things would have been different were some of the conditions different. The contrastive counterfactual theory of causal explanation can illuminate this idea, too (Ylikoski 2009).

[28] Notice that this is not a *reductive counterfactual theory of causation* aimed at defining causation with non-causal concepts (e.g. Lewis 1973; Vihvelin 1995): it presupposes the concept of causation as a primitive. The analysis is meant to be a tool for assessing causal relations more complex than a primitive cause–effect relation. (See Ylikoski 2001.)

make a commitment to any specific ontology of causation.[29] Furthermore, it is not an attempt to give criteria for *the* cause of an event. It is an

[29] There is a *de facto* pluralism of causation concepts that probably confuses a lot of discussion on causal explanation (see Sober 1985; Hitchcock 2001; Cartwright 2004; Hall 2004; Godfrey-Smith 2010). For the present purposes, it is not necessary to participate in the debate on whether there is a "real" concept of cause and what it is if it exists. But it is important to clarify how the notion is used here – what the aim of causal explanation is taken to be. First of all, there are at least three conceptually different objects for a theory of causation: (1) the *fundamental metaphysical* issue of what connects events across time and is presupposed for the laws of physics to exist (but is not studied by physics); (2) the *pragmatic notion* of causation that refers to things being connected in a certain way without this relation belonging to fundamental ontology (and perhaps not to physics either; see Price & Corry 2007 for causation in physics, von Wright 1971 and Woodward 2003a for the pragmatic notion of causation); and (3) the *physics* of causation that explains how the physical structure of the world is related to pragmatic-level causal relations (for example, Salmon 1984 and Dowe 2000). There can be different, contradictory views on how these issues are related and whether any of them are pseudo-issues (for example, (1) is a pseudo-question and the answer to (3) gives the most fundamental level of causation; (2) must be reduced to (3); or (2) is the only true sense of the concept). The notion of causation used here is the pragmatic notion. Secondly, the pragmatic notion can be taken as a *difference-making* relation or a *productive* relation (Hall 2004; Godfrey-Smith 2010). These two notions imply two different criteria of validity for causal explanations. The difference-making relation is the prime focus here. It is the most "minimal" notion of causation and the one most broadly applicable to causal explanation in practice. The other notions include some extra criteria that tend to be metaphysical in nature. If the minimal notion of causality is not satisfactory because it does not include any *ontology* of causation, we can simply consider the term "causation" to be shorthand for *causally explanatory relations*. This does not change the substance of the discussion. The same answer applies to causal pluralism in the other sense, one that is implied by the contrastive-counterfactual account: we can give the same event several different causal explanations that cannot be thought of as combinable partial causes. This is the case with natural selection as well as mental causation, both of which will be discussed later and in which the ontological attitude seems to lead into eliminativism of an important explanatory factor (see also Shapiro & Sober 2007).

account that explicates a notion of cause that is applicable in different contexts in which we need to talk about causal dependencies, including partial and negative causes as well as causes on different levels (Hitchcock 2003). Even if causation necessarily involves some continuous physical processes directed by strict laws of physics (Salmon 1984), complex causal systems may have nets of causal dependencies through somewhat constant structures (that we do not need to pay attention to under normal circumstances), or a *causal field* (Mackie 1974) in which both the cause and the effect are changes. It may be enough for a causal explanation simply to know the existence of the connection, not the details of how the connection is constituted. Biological explanations typically take place in systems or parts of systems like this, and it almost never makes sense to go all the way down to basic physics in the explanation.

All this has several consequences that are will be important later. First, as mentioned above, things can be causally dependent even if they are seemingly unconnected and would not be in a causal relation without the other factors within the system. Secondly, negative causes (omissions, preventions, and breaks in the system) can be treated as causally effective events when the system itself, functioning properly, is considered a background condition. Both positive and negative causes require a context; within a context, they can be treated symmetrically in explanations. Thirdly, it is sensible to refer to a continuous causal structure with descriptions that are on different "levels" of description (for example, macro and micro levels in social sciences) and yet identify causes and effects correctly. In the last case, a "higher-level" description accurately refers to a type of conditions that are realized one way or another on the "lower level." But in some cases, the higher-level description, which might be functional and indirect, may even be a more precise way to extract the causally relevant factors. For example, if the description refers to a robust function within the system that can be multiply realized, it may be more informative to use that function as an explanatory category and leave the detailed

description of the instantiation as a black box.[30] What matters for the adequacy of an explanation is that it correctly identifies the difference-maker. The fundamentality of the level of description (whatever that means) does not. The difference-maker may be a relational property or a cluster of properties on a lower level and captured as an abstract, higher-level property in the explanatory description. The contrastive-counterfactual theory provides the logical form of an explanation that allows us to account for these aspects of a complex causal system by giving criteria for the constrained, unequivocal individuation of both the explanatory cause and the explained effect (see Ylikoski 2001; Woodward 2003a). What *kinds* of things are causally dependent in a relevant way is an empirical issue?

The counterfactual theory can be further refined with a *manipulation* or *intervention* account of causation: the idea that *manipulation* is central to the idea of causation (von Wright 1971; Menzies & Price 1993; Pearl 2000; Woodward 2003a & 2003b). According to this view, having a causal connection between two events means that if we could *intervene* in the process by preventing or producing the cause, we would also prevent or produce the effect.[31] *A* causing *B* means that if

---

[30] This is also roughly the idea behind the increasingly popular stance on mental causation advocated by Menzies 2007, Shapiro & Sober 2007, Woodward 2008, and Raatikainen 2010. According to them, causal claims under mental description are not competitive with or additional to causal claims about the same events under physical (that is, neural) descriptions. On the contrary, they are two different ways to refer to the same events (changes in the system), but a mental description may be a more accurate way to pick out the difference maker for some purposes.

[31] This intervention is ideal: we would not, in many real cases, be able to prevent or produce just one event or aspect of a system. But it does not matter for the logic of the causal explanation. Even if the system includes factors with multiple causal functions, loops, and feedbacks, as many biological systems do, the causal roles that make up the system (or a mechanism) and the actual parts that instantiate these roles are *analytically* separate. The actual interventions would be impossible since affecting one function necessarily affects

we prevented *A* we would also prevent *B*; and that if *B* had not been the case before the intervention, if we then produced *A* (under the same background conditions), we would also produce *B*. Our knowledge of causal relations is knowledge of what we could manipulate in order to change the outcome. This explication is not a reductive explication of the concept of causation, for it uses causative notions like "produce" and "prevent" that presuppose causation, but it explicates some important aspects of causal explanation as a scientific practice. First, it enables us to formalize causal relations in a complex system (such as any biological developmental system) that include overdetermination (for example, multiple alternative causal routes between a particular partial cause and effect in a canalized developmental processes) and "exotic" causal relations like preventions of preventions (for example, a gene blocking another gene blocking a developmental path, therefore enabling the developmental path) (see Pearl 2000; Woodward 2003a). Second, the idea explicates the rationale of experimental practices of discovering causal relations (Woodward 2003b).

Third, the interventionist idea explicates a pragmatic and more limited interest for having causal explanations: we are sometimes interested in knowing which factors in the system contribute causally to the *explanandum* in a particular way, such that changing them would produce the right kind of effect from the perspective of our explanatory interests. Sometimes, however, this definitional relationship between causes and effects can become reversed because of the explanatory resources available. For example, if we are interested in the effects that hormonal changes have on some behaviour (such as sexuality or aggression), the *explanans* (hormones) redefines the *explanandum* accordingly, which means that we are no longer explaining sexuality or aggression, but a certain effect within them that is explainable by the *explanans* of choice (see Longino 2013).

---

some other functions in the system at the same time, but the logic of what it means to be a cause according to the interventionist account does not break.

Not all interventions are relevant to all our explanatory interests. Some causes are kept constant as the causal background for current explanatory purposes. This is not to say that they are not causally relevant, but that we are interested only in some other changes. We may be interested only in the range of factors that we can actually intervene in (in contrast to ideal interventions), for example. In the case of historical explanations (such as evolutionary explanations), we cannot perform actual interventions, but some alternative courses of events (that is, theoretical interventions in the evolutionary history) are more plausible than others and we might want to consider only those. Some other interventions, in turn, would have too large an effect on the whole system to be meaningfully treated as a cause for the specific effect that we are attempting to explain. For example, to return to the example of cat aggression, a genetic change that blinds or paralyzes the cat, or prevents the aggressive behaviour in general, would also prevent the aggressive response to a small moving object, but we would not consider pointing out those difference makers as explanatory, or those genes to be genes "for" that behaviour. Although the presence of those particular genetic resources is a necessary causal factor for the behaviour in interest to develop, it is not *explanatorily adequate*. The contrast presupposed by the causal explanation is more precise. More generally, if a gene is understood as a DNA sequence, there is no gene that would be a gene for any phenotypic trait as such without being a part of a complex net of causes. All talk about genes "for" a trait presumes a fixed set of certain background conditions that may include genetic, epigenetic, and environmental factors, and a change in the background may break the causal connection.

In other words, causal factors are *explanatorily informative* depending on both the real causal relations in the world and the explanatory interests we have. An explanation seeks a difference-maker between possible outcomes that are alternative to what actually happened and mutually exclusive within the contrast class that is presupposed by the explanatory question – it does not seek any or every causal factor that the *explanandum* is dependent on. For example, the

explanatory question about the cat exhibiting aggressive behaviour towards a small moving object asks for a difference maker between this behaviour and the alternatives, including the cat not being interested in the object or having another kind of response. However, the cat would still have to be able to perceive it and to act in the aggressive way in principle – it just would not be inclined to do so. If an alternative implied by the difference maker would remove the disposition to be aggressive, to perceive the movement, or the ability to act, what is being explained is something else than what we are interested in explaining.

There are also other ways in which all causally contributing factors are not equal from the point of view of explanation. First, from a pragmatic point of view, differences in explanatory interests may arrange the causal factors in different orders of importance. For example, whether we should consider genes more important explanatory factors in development than equally necessary environmental factors, or whether we should give adaptation a special role when other evolutionary factors are in play as well does not depend solely on the facts of the matter. Secondly, depending on the nature of the causal network we are dealing with, there may be further differences between causal relations. For example, there may be differences in the following aspects (see Woodward 2003a & 2010 and especially Ylikoski & Kuorikoski 2010):

1) how **stable** or **non-sensitive** they are: how much change there can be in the background conditions for the explanatory relation to hold;

2) how **precisely** the description captures the relevant causal information: how much relevant detail the description contains or how much irrelevant detail it omits – an important case of this is the *graining problem* and the choice of an appropriate level of abstraction in description; and

3) how **specific** the link between the cause and the effect we are interested in is: how exclusively is the cause related to the particular effect in contrast to having a lot of effects in the system.

There may be trade-offs between these properties in choosing the explanation. Furthermore, if we are interested in how the cause and effect are connected, for instance by explicating a mechanistic connection (I will return to this soon), a further property of an explanation is the **accuracy** of the details of the causal connection. This may be another factor in the trade-offs.

2.1.4. Invariance in Explanation

Explanations require a link between the *explanandum* and *explanans*. The fundamental theories of explanation aim to explicate what constitutes this relation in the world. The contrastive-counterfactual theory does not take a stance on the issue but only presupposes dependencies. Discovering them is left to empirical sciences. In practice, sciences discover generalizable relations between types of events, properties, and so on. From the point of view of the contrastive-counterfactual theory, what matters for these generalizations to be explanatory is *invariance* (see Woodward 2000, 2001 & 2003a): the generalization about the *relation* between two types of, say, events, continues to hold under an intervention of a kind described above. However, it does not need to hold under any intervention to the background conditions that are presupposed by the generalization.

For example, we may want to fix the environmental factors that are reliably present within the whole range of a species' natural habitats, as well as the shared genes, as the causal background in the development of the members of the species for practical purposes and concentrate only on the factors that *actually* cause individual differences. Furthermore, even if we are interested in the developmental processes with fewer arbitrary background assumptions like this,

some factors (genetic or environmental) are necessary for the development but do not make a difference in the outcome, while others do. We can use the contrastive theory to distinguish between *instructive factors* (the *specific* environmental or genetic factors that are responsible for individual differences in developmental outcomes) and *permissive factors* (the environmental factors and genes that participate in the process but can be black-boxed for the purposes of explaining with these change-relating generalization; see Woodward 2010; Griffiths & Stotz 2013; Calcott 2017). This enables us to talk about a gene *for* a trait even when the development of a trait is dependent on various other causal factors to which we do not pay attention. If the environment or population's genetic structure changes, however, the scope of relevant explanatory factors may change. Most generalizations have a limited range of changes over which the relation holds, and they break under extreme interventions. For example, if a generalization is about quantities (for example, an increase in the exposure to some substance increases the risk of getting a certain disease), there is usually a range of values under which it holds, while it breaks when the values exceed this range. Woodward (2000) calls this range of changes under which the generalization holds the *invariance domain*. Presumably, all non-physical generalizations (and most physical generalizations, for that matter) have an invariance domain.

For many pragmatic considerations, we do not need to know the extent of the invariance domain and its conditions exactly. It is sufficient to know whether the range of variation that is important for us for practical reasons (the natural habitats, for example) falls within the invariance domain. From the perspective of the basic research, however, what interests us most is precisely the contribution of these "normal" conditions. The relevant range of variation in background conditions may also vary from discipline to discipline. For example, a psychologist and a biologist may hold different environmental factors to be constant in development. Consequently, some key concepts may have broader reference in some approaches than in others. I will later argue that this is the case with the concept of innateness: it may have

a practical use (and a reference) from the psychological point of view in some contexts, while it is misleading at best from a biological point of view.[32]

The limits of an invariance domain are a reason to be interested in the basis of the invariance, not just its existence (for practical purposes): why and how the cause and the effect are related in an invariant causal relation. We might also be interested in this for further scientific understanding of the phenomenon, or for purposes of discovery, for example. A strict natural law (of physics) would provide such a connection as a primitive fact of matter, a characteristic of reality. However, regardless of our stance on such fundamental laws, biological systems are complex systems in which *further* invariances take place in functionally organized structures. Although they do not break any laws of physics, they exhibit regularities in behaviour that are not necessitated by the laws of physics alone but also by how the systems are structured and what are the functional relations between the parts. There are lawlike biological generalizations, too, but they are notoriously contingent, local, and fragile (Beatty 1995; Glennan 1996 & 2002b; Mitchell 1997 & 2000; Raerinne 2011). Although these generalizations can be used in explanations and predictions under the contrastive-counterfactual model (see Woodward 2001; Raerinne 2011) and in modelling the behaviour of a system, there is a call for further explanation of their very existence, too (see also Andersen 2011). Moreover, the basis of a generalization enables us to make predictions about the range of applicability, or the invariance domain. The idea of *biological mechanisms* has been used to accomplish this, among other things (Bechtel & Richardson 1993; Bechtel & Abrahamsen 2005; Darden 2006; Glennan 1996, 2002b & 2010; Machamer, Darden & Craver 2000; Craver 2007).

---

[32] There is also a vernacular notion of innateness that is a part of folk biology and folk psychology and refers to a completely invariant appearance of a trait, which is also the notion of innateness traditionally used in philosophical discussion of "innate ideas". Nothing is innate in this sense. (See Sober 1998; Medin & Atran 1999; Griffiths 2002; Bering 2006; Bateson & Mameli 2007.)

2.2. Biological Mechanisms

The idea of (non-physical) science as a search for mechanisms, not laws, or the *New Mechanistic Philosophy*, is a general idea of how to understand scientific discovery, explanation, prediction, and modelling. It has been used to account for things like causation, levels, function, reduction, and emergence in complex biological systems and beyond. (See Machamer, Darden & Craver 2000; Bechtel 2006; Darden 2006; Craver 2007; Wimsatt 2007; Glennan 2010a; Levy 2013.) There are two basic ideas in particular that are relevant to what will follow. First, the idea that knowledge about mechanisms provides explanatory information on how and when a causal connection holds between two things of an interest. Opening the black box tells us when the black box can be kept closed for practical purposes, if the mechanistic basis is known to hold in all the practically relevant contexts. Secondly, mechanistic thinking may help to relate different *kinds* of causes (and therefore different kinds of causal explanations) that constitute a mechanism from the perspective of the functioning of a bigger whole. I will use this line of thinking to articulate the relationship between proximate, developmental, evolutionary historical and evolutionary functional explanations. Before this, I will discuss the mechanistic approach in biology in general and in understanding natural selection explanation in particular.

2.2.1. Mechanisms in Explanation

The idea of mechanistic connections being an important part of understanding biological phenomena goes back to the Early Modern era, although it was abandoned for a while (see Grene & Depew 2004). The idea of mechanistic explanation was introduced to more recent discussion by Peter Railton (1978) and Wesley Salmon (1984). They thought of mechanisms as *interacting causal processes*. Contemporary mechanists, building largely upon William Wimsatt's pioneering

work (see Wimsatt 1974 & 2007), define mechanism as a *system of interacting parts*. The seminal paper by Peter Machamer, Lindley Darden and Carl Craver, for example, defines mechanisms as "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions." (Machamer, Darden & Craver 2000: 3.) The earlier characterizations concentrated on the basic idea of mechanism as *interactions of partial causes* that result in *emergent causal connections* in a complicated network of causal processes, whereas the contemporary discussion concentrates on the functional structures of such causal connections. The more novel approach, the New Mechanistic Philosophy, attempts to provide insight on the mechanistic connections in nature precisely through an analysis of that structure. However, the difference between mechanism as *process* and mechanism as *structure* is crucial (Glennan 2010b), and it will have consequences for how to think about different kinds of mechanisms. The idea of mechanistic interaction of processes may have a broader scope than the structural approach.

I will not go into the various definitions of mechanism and the disagreements in the literature, however. I will only describe the general idea of mechanistic explanation briefly and point out that there are different ways in which the basic explanatory logic may be implemented. Much of the discussion on biological mechanisms is about physical structures such as a cell (Bechtel 2006) or brain (2007), but social mechanisms, for example, work quite differently: the parts of the mechanism are types of interaction (see Kuorikoski 2009). There is no reason why some biological systems would not work in a similar way – for instance, social behavioural traits that are constituted by interactions. I will use this idea precisely in that context. Furthermore, as we will see shortly, natural selection itself is a mechanism of the latter kind. But before getting to these issues, I will outline the general logic of mechanistic explanation that I assume to be generally accepted by the mechanists.

The general idea of mechanistic explanation is the following (abstracting from Glennan 1996, 2002b & 2010b; Machamer, Darden &

Craver 2000; Bechtel 2006 & 2011; Craver 2007 & 2103; my paraphrasing). The *mechanistic structure* is a system of entities and activities with a functional structure (such as the composition of a cell or a social structure that characterizes the relations between people). The parts of the system work in an orchestrated manner in a causally functional, invariant relation to each other. When the causal properties of the parts are combined with the structure of the interaction between the parts, a causal pathway emerges, making a causal connection, a *mechanistic process* between two entities or events not connected otherwise. This gives the entity or event thus created a *mechanistic causal explanation*. The regularities in the mechanistic structure entail generalizable explanatory regularities. Furthermore, a mechanistic structure can give rise to a new phenomenon and properties that are of the mechanistic whole and constituted by the parts and their activities. This is a *mechanistic constitutive explanation*.[33] These regularities, phenomena and properties would not exist because of the parts alone, but also need the *structure of invariant relations* to exist. The architecture of the mechanism has a constitutive role in mediating the causal powers or regular behaviours (whichever perspective is chosen for the analysis) of the parts, transforming it into the causal power or regular behaviour of the whole system. The causal powers or regular behaviours of the component parts are in turn constituted by lower-level mechanisms that are guided by further lower-level mechanisms, and so on, until the fundamental regularities are reached.

The "levels" in mechanistic explanations do not refer to *ontological* relations[34] or disciplinary relations. They refer to the hierarchies of part-whole interactions in which entities of the same type (for example, molecules) may be parts of a mechanism on different levels of

[33] For the difference between causal and constitutive explanations by mechanisms, see Salmon 1984; Bechtel 2006; Craver 2007; Kuorikoski 2012; and Krickel 2018.

[34] Until, perhaps, at the explanatory "bottom," if physicalism is correct and causation can be given a physical definition, both of which are contested issues, and neither of which matters in the context of this dissertation.

hierarchy (for example, constituting a neuron, which in turn is a part of a neural process, and participates directly in brain processes; Craver 2007). What is important is the *causal functional structure* and not, for example, physical hierarchical structure as such. Nevertheless, a key insight in the New Mechanistic Philosophy is that the physical structure and causal functional structure can be studied together in a dialectical manner.

However, mechanistic systems function in an environment. The environment is not just a source of causal inputs and a target of outputs. A fixed system in a causal interaction with its environment may be structured so that certain parts of the environment, when present, are constitutive parts of a causal *process*, although they are not a part of the system in a narrower sense.[35] These external factors may be considered part of the mechanism as a process bringing about a change in the states of the system, even if the explanatory mechanistic structure is conceived to be the fixed structure alone. Such interactive processing is a crucial part of many biological systems that use resources reliably available in the environment – not only in their operation and maintenance of homeostasis, for example, but also in their development (Wolpert *et al* 2010).

The general idea of a causal mechanism as systemic interaction of causal parts could be used as a theory of causation, a theory of explanation, or a theory of modelling strategies (see Levy 2013). I am not using it directly as any of the above. I use it as a supplement to the contrastive theory of causal explanation, as a framework to characterize the processes and structures in the world that constitute complex dependencies. Not all explaining is mechanism-seeking, not even when there are mechanisms involved. On the contrary, mechanism-seeking explanations presuppose the idea that the mechanism is a

[35] One may be, of course, either internalist or externalist when it comes to characterizing a mechanistic system such as a cell, but there may be both pragmatic and theoretical explanatory reasons to draw a clear distinction between a physical structure such as a cell and its environment.

mechanism *for something being the case* instead of *something else being the case* (Craver 2013). Explaining with a mechanism, or rather, with a *mechanistic connection*, presupposes some idea of why we are interested in the mechanism for explanatory purposes in the first place, and the contrastive-counterfactual theory can be used to explicate this. This is important with complex systems in which the structures and processes may have multiple functions and can be partitioned in different ways so that different combinations of parts constitute different mechanisms. A mechanism discovered for a behavioural trait, for example, is not likely to be *the* mechanism for this behaviour (see Longino 2013). All mechanistic explanations are answers to more precise explanatory questions than that. The individuation of a mechanism involves deciding what effect we are trying to find a mechanism *for*. If we are not supplementing mechanistic approach with a contrastive-counterfactual or some other normative theory of explanation in this fashion, we would end up just describing structures of causal interactions without criteria for individuation of *explanans* for any given *explanandum*. This would probably be simply fine, or even desirable, for a fundamentalist explanatory project seeking a complete set of factors. Given the practical needs for explanatory theories that are under consideration here, that would not do. This is important, for example, for understanding natural selection, as we will see.

Discovering mechanistic detail may deepen the explanation in three ways. First, in providing an explanation for a *singular explandum*, adding mechanistic detail helps to make it more *precise*, more *specific*, or both. Second, in providing an explanation for a *generic explanandum*, knowing mechanistic detail tells us about the invariance domain and the *stability*. It provides explanatory information about the conditions under which a causal connection holds between two things of interest and when it breaks, as well as an insight into abnormal functioning and its causes. Third, our knowledge of mechanistic detail provides deeper understanding of the *explanandum* in the sense of increasing the explanatory counterfactual information: we are able to answer quantitatively more and qualitatively different kinds of what-if-

things-were-different questions (Ylikoski & Kuorikoski 2010). Moreover, mechanistic understanding provides a way to evaluate the connection between different (but correct) explanations by locating them in a wider scheme of mechanistic connections. This will be my general understanding of how the different kinds of explanation (including the three different explanatory dimensions) are related in the evolutionary explanations of behaviour.

Our knowledge of the existence of a connecting mechanism, even when it is left black-boxed, is evidence for either a generic or a singular causal claim between cause and effect, or both.[36] Our knowledge of the details of the mechanism tells us what makes the connection and what the conditions for the causal relation to hold are: the structure of the mechanism and the causal functions the parts play in the whole system. Detailing a mechanism opens the explanatory black box. However, given that the mechanistic thinking itself presupposes both simpler causal relations (of the component parts) and needs an account for individuation, mechanistic thinking is not a stand-alone theory of causation or causal explanation. Whether or not a mechanistic description is needed or helpful is a case-by-case issue that depends on purposes. (Ylikoski 2001; Woodward 2002 & 2011.) But if the essence of mechanistic thinking in explanatory contexts is to articulate the connection involved in the causal relation between the *explanandum* and *explanans*, and that the *kind* of details used for describing a mechanism has no relevance as such, we can be liberal as to what

[36] It is possible that there are unique mechanisms in the sense that the exact architecture is instantiated only once. A unique mechanism like this (for example, a unique historical situation combining causally interactive factors in a novel way) could create a unique causal connection between events that holds only once. This means that there can be singular causal relations that are not instantiations of any causal generalization. (See Glennan 2002b.) However, this does not mean that there are no underlying generalizable causal processes or sub-mechanisms (as component parts), and from an epistemic point of view, we can know about the causal nature of these individual cases only through our knowledge of these underlying processes or mechanisms.

counts as a mechanism. In fact, we must allow a variety of different kinds of mechanisms if we want the mechanistic thinking to be useful in all the complex causal structures that it is intended to be useful for.

An important link between manipulationist and mechanistic thinking about causation has to do with the causal links *within* the mechanism. A description of the causal connections between different parts is needed to characterize the internal workings of a mechanism. A manipulation account of causation can be used for this (see Menzies 2012). There is, however, a complication when many causal parts of a complex system have several causal functions, there are looping effects, and so on. This has been used as an argument against the manipulation notion of causality in this context, and an alternative view has been provided, according to which the mechanistic explanation is simply about describing the structures and the activities of the parts of the mechanistic system (Machamer 2004; Bogen 2005). The core of the criticism is that one cannot make isolated causal interventions in the system, and the manipulationist notion, according to the critics, seems to require this. This criticism is, however, not valid. The idea of manipulationist analysis is to conceptualize the causal relations in a piece-meal fashion that distinguishes between the *different causal roles* a part plays in the system. The notion of an ideal intervention is used to articulate this idea in cases where no actual intervention would be conceivable without affecting the system otherwise. This disconnects the notion from the actual manipulation, but the aim is to describe the functional structure of the causal networks, not the physical structure that instantiates them. The *causal functional decomposition* needs not be identical to the *structural decomposition* of the system. The same parts of a concrete mechanistic structure can instantiate several different parts of the causal system that is instantiated by the structure. Therefore, if one simply describes the structure, the activities of the parts, and the processes flowing through them, some causal information about the system is *missing*.

Not all mechanisms are describable as processes and interactions between concrete parts of a fixed structure, yet the mechanistic

explanatory logic seems to apply. This is the case with *social mechanisms*, and there is no reason to think that there could not be mechanisms like this in biology, too. As I mentioned at the beginning of this section, I will argue that social behaviour should be understood within the alternative framework even from the biological point of view, and I will apply this idea to natural selection as well. In a social mechanism, the same concrete "parts" (individuals) may play different causal roles in different contexts, and these functions, not the individuals, are relevant to the explanation. These social mechanisms can still be described as systematic interactions between causal processes that give rise to constitutive and causal relations that would not exist without the structure of causal interactions (Kuorikoski 2009). Robert Skipper and Roberta Milstein (2005), in turn, have argued that the standard models of mechanisms fail to describe natural selection as a mechanism. This is because, according to them, natural selection is a probabilistic process and its component parts (its entities and activities) cannot be individuals, traits, or the population itself. They are mostly right, but this criticism only states that natural selection cannot be a *kind* of mechanism that is characterized in *this* way.[37]

Jaakko Kuorikoski (2009) has argued that there are two different concepts of mechanism that play the same explanatory role but are different in important ways: mechanism as a **componential causal system (CCS)** and mechanism as an **abstract form of interaction (AFI)**. The CCS mechanism is the one that has been discussed by the New Mechanists, but there is no reason to consider only one of them "really" a mechanism. Both of these notions are referred to as "mechanisms" by scientists themselves and both are based on the idea that some causal connections are the results of nets of causal processes that have a structure that grounds the emerging explanatory connections.

[37] The part of their criticism that is not correct is the probabilistic part. For a reply to the problem of probabilistic processes in general, see Barros 2008; for a defence of the possibility of stochastic mechanisms, see DesAutels 2011. This problem does not arise the same way in the notion of mechanism I am moving into now, so it will not be discussed here.

The explanatory logic is similar enough, attributing the existence of causal generalizations to robust effects of networks of causal processes instantiated by the constituent parts.[38] The nature of the parts and their interaction may be different with different kinds of mechanisms. An adequate articulation of mechanistic thinking should include all types of mechanistic explanatory strategies, if possible, and then proceed to the finer points of different mechanisms if needed. Therefore, as a starting point at least, both CCS and AFI are concepts of mechanism at minimum, or, as I would rather say, they are sub-categories of the same explanatory idea and different, richer elaborations of a more minimalistic concept of mechanism.[39] Both are important explanatory concepts for our current purposes.

---

[38] The New Mechanistic Philosophy starts with discussion about CCS mechanisms and that has been the subsequent emphasis. However, it is a contingent fact that the philosophers of science started to analyse one category of mechanistic connections under the banner of mechanism. Surely, this is not an adequate reason to restrict the notion to these cases only, especially when the term is used more extensively in actual scientific practices.

[39] Stuart Glennan (2015: 145) defines a *minimal mechanism* as follows: "A mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized so as to be responsible for the phenomenon." This is sufficient to characterize the logic of mechanistic explanation, but any given context of explanation requires some analysis of how the parts achieve this or evidence that they do. I consider the more sophisticated definitions to be explications of mechanistic logic under more specific conditions. The different definitions for a mechanism (for example, Bechtel & Richardson 1993; Bechtel & Abrahamsen 2005; Glennan 1996, 2002b & 2010; Machamer, Darden & Craver 2000; Craver 2007) do not need to be competing definitions for the same idea of mechanism but specify different conditions that have different consequences and apply in different contexts, granting contextual rather than universal insights about mechanisms. For example, such important issues as explaining the distinctive capacity of living things to maintain their homeostasis as systems separate from their environment (which is a part of being alive; see Ganti 2003 and Bedau 2010 for more detailed analysis) require additional properties of the mechanistic explanatory models (Bechtel 2011).

CCS mechanisms are Cartesian mechanisms, physical systems (such as a cell in Bechtel 2006 or a brain in Craver 2007) that are both functionally and structurally decomposable into concrete component parts. There are specific, direct causal connection between the parts, and these are the basis for the macro-properties of the system due to their causal powers and the composition of the system (see also Machamer 2004; Glennan 2010a). In AFI mechanisms, in contrast, the component parts of the mechanism are the *kinds of behaviour* that can be instantiated by various concrete parts that, in turn, can potentially instantiate multiple behaviours. For example, in studying social mechanisms, the concrete component parts (individuals and their behavioural dispositions) are interchangeable with each other in a social context, and the same individual exhibits different causal capacities in different contexts. The individuals are not only related to each other in a structure, but their relevant causal properties are *relational:* they depend on other individuals and the context. This context-dependent behaviour and its function in the structure of interactions on the population level is what matters, not particular individuals and their individual relation to each other, and these forms of interaction can be abstracted from the concrete interactions as the constitute parts of the mechanisms. The causal functional structure cannot be mapped to a spatiotemporal structure made of individuals, but to the patterns of interaction, that in turn are dependent on a repertoire of behavioural choices of individuals and how they manifest in the interactions.[40]

[40] There is a direct link from this to the individualism–holism issue. The agent-based models may model either individualistic or holistic processes. This is because the agents in the models may represent fixed individuals or individuals that play a causal role (which may or may not be a social role) that is specified by the context and affected by other factors in it. As Kuorikoski (2009: 35, n35) points out, the AFI approach identifies causal properties that are, as Wimsatt (2007: 217) puts it, "in between levels." The account that I will give of some of the social behavioural traits as causally explanatory, fitness-difference-making factors in the evolution of social behaviour, can be characterized in this way.

Since the causal powers of the "parts" (the individuals) depend on other parts directly, the system is not a structural composite of components with intrinsic causal powers. Instead, the dynamics of the system must be understood by abstracting the *forms* of interaction within the system (such as interactions between individuals). However, if we adopt the interventionist conception of causality and the contrastive theory of explanation, we can operate with mechanistic causal functional structures abstracted from a network of causal processes.

Arnon Levy and William Bechtel (2013), in turn, discuss *abstract mechanisms*, by which they understand patterns of causal connectivity (such as network models and causal graphs) that are abstractions from the details of the mechanistic basis for those connections. In contrast to some other mechanists (for example, Machamer, Darden & Craver 2000; Darden 2006; Craver 2007), they think that when it comes to modelling mechanisms, there should be room alongside the detailed models of decomposable parts and activities for the idea of abstract models of the behaviour of the system that have an explanatory role. The general point is that it is not always necessary to go into the details of concrete interactions of the parts of the mechanistic structure. It could even be misleading.

I will discuss behavioural traits and psychological traits in social interaction as two different (yet connected) explananda for evolutionary explanation in a later chapter. It is important to distinguish individual responses in a social setting and *specific* intentions or psychological tendencies – for instance, altruistic behavioural tendencies and altruistic motivation. The behavioural function in any given social setting is multiply realizable by various psychological factors. Referring to psychological states instead of behavioural interactions would be more precise but not robust enough to have maximal explanatory information for some explanatory aims if the underlying psychology is heterogenous on the population level. Furthermore, if we abandon *a priori* individualism in our attempt to describe the components of social interaction, we may want to go *deeper* into the cognitive and motivational architectures of individuals instead of staying at the

intentional-level description. The sub-personal cognitive processes and motivational mechanisms that constitute individuals as agents, as well as their behavioural dispositions, may also *directly* constitute supra-individual phenomena by being parts of social mechanisms in contextual settings (see Sperber 1997).

2.2.2. *Natural Selection as a Mechanism*

Biological evolution is a population-level change in the frequencies of individual properties, usually measured by genetic variation underlying the phenotypic variation. This is not conceptually necessary and there are good reasons to think that there are alternatives to genetic inheritance, but I will keep the idea simple for now. There are many sources for the change. Firstly, there could be a source of novelty, such as mutation. Secondly, given that all individuals do not produce exactly the same number of further reproducing offspring, there will always be effects of chance and even disappearance of variation for purely statistical reasons – this is *drift*. In principle, these factors alone, combined with reproductive isolation keeping the mutations and drift apart, could even lead into speciation. However, there can be systematic biases caused by the properties of the evolving traits themselves in this chance process, such as their usefulness in the environment that the population is spread across. This requires a combination of causal factors that together form the *mechanism* of *natural selection*. Natural selection can be defined, following John Endler (1986), as a process in which

C1) there is *variation* in a given phenotypic trait on the population level,

C2) there is a consistent relationship between the trait and *fitness*,

C3) the trait is *inherited*, and because of this,

O) there will be a change in the frequency of distribution of the variation in the trait between generations (see also Lewontin 1970; Futuyma 1998).

This definition describes the logic of a causal mechanism: given the component parts (C1–C3) and their causal contribution to the system, an outcome (O) is produced. The causal processes go through the individuals and their interaction with the environment, and the reproductive system, but the relevant causal components in the explanation are C1–C3. The generalizable phenotypic differences between the individuals (C1) are abstracted on the population level (that is, they are differences between phenotypes, not individuals), and although the differences are not causal in the *productive* sense, their existence is a necessary difference-maker in the mechanistic structure that produces the change. C2 refers to the systematic bias these phenotypic differences cause in the expected reproductive success (which is a relational property, in respect both to the other individuals of the population and to the environment).[41] C3 expresses the presupposition of a mechanism that transfers the trait to the next generation in the population. These factors causally direct the process of evolution. The outcome is a population-level change in the variation to a certain direction. This process cannot be understood as a Cartesian mechanism, but the aspects of the real interactions that the individuals within the population have with their environments, that are causally relevant to C1 and C2, can be abstracted as the causal components interacting with each other and C3, producing O.

Although the concrete causal *processes* take place at the level of organisms and their interaction with their environments, some aspects of environment and individual characters have systematic

[41] Note that the reference to fitness is not a reference to a cause for evolution here, but a reference to an effect that the phenotypic differences have on reproduction. I will discuss the concept of fitness more extensively later, when I discuss the concept of evolutionary altruism.

effects that can be abstracted and generalized on the population level. These abstracted aspects can be considered components of causal relations in the framework of the contrastive-counterfactual theory of causal explanation. Put together, these components constitute a *functional mechanistic structure*. The component part C3 is a CCS mechanism that is presupposed but often black-boxed in evolutionary explanations – I come back to this later. The mechanistic nature of natural selection is both a model example and a test case for New Mechanistic Thought. It is a model example because the mechanistic nature of natural selection explains its productivity as well as the observable regularities without laws, and the mechanistic nature makes its inefficiencies and fragility understandable. It is a test case because applying the most common definitions of a mechanism has proven to be problematic. I submit that the process of evolution by natural selection and its causal factors cannot be adequately presented as a CCS mechanism, but the mechanistic logic on an AFI works just fine. Since the topic of this dissertation is evolutionary explanation and it is approached from the mechanistic perspective, I will briefly discuss this issue and the lessons from the debate.

One key issue in the debate on natural selection as a mechanism has to do with the fact that its mechanistic components are abstractions (Skipper & Millstein 2005). Some have questioned whether it can be a causal factor at all. The advocates of a non-causalist, *statisticalist* alternative (Walsh 2000; Walsh, Lewens & Ariew 2002; Matthen & Ariew 2002 & 2009) claim that natural selection is merely a statistical product of the causal interactions of individuals with their environments. Following from that, the real causal factors are the same that are at work in drift. According to them, the difference between natural selection and drift is just a difference in the statistical *outcome* of the causal processes of evolution and neither of them are causal factors themselves. This approach, however, focuses attention on all the wrong places. The critical issue seems to be that they approach causation and causal explanation from the fundamentalist point of view, tracing singular, concrete causal processes: for them, causal factors of

evolution are things such as predation, sunlight, and direct (in contrast to evolutionary) competition (Walsh, Lewens & Ariew 2002). This particular problem does not arise in the framework of the manipulationist conception of causation and the contrastive-counterfactual theory of causal explanation (see also Reisman & Forber 2005; Shapiro & Sober 2007). Furthermore, if natural selection is an AFI mechanism, its explanatory mechanistic structure does not need coincide with the actual causal processes. The reasoning behind this follows.

"Natural selection" as such refers to the abstract form of interaction in which the component parts are whichever concrete factors have the relevant causal functions specified by the description of natural selection – all particular explanations "fill" these "forms" (or "recipes", to follow Peter Godfrey-Smith's 2009 terminology) with traits and environmental factors. It may be correct to say that natural selection, in abstract, is not a universal mechanism – but then again, it does not even exist as such. It is not a natural law, but a description of a certain functional structure of a causal system – any causal system. This functional structure can be instantiated by concrete systems of causal interaction in which the functioning of the concrete parts constitutes local AFI mechanisms through the network of causal processes. The same logical structure can be found in various non-biological contexts, too, such as cultural evolution. It is also important to notice that both the causes (abstracted inheritable differences measurable in fitness) and the effects (changes in the frequencies on the population level across the generations) occur at the population level. Even if the *causal processes* take place on individual level interactions, the *causal explanation* "slices" the entirety of the causal interactions on the population level with a description that captures those properties of the concrete interactions that instantiate the logic of natural selection. At the same time, if the theory of natural selection applies to a particular change in a particular population, then these properties are causally efficient in biasing the evolutionary process.

There are various descriptions that may refer to the causal factors that *make the difference* in the outcome, but the population-level

description that uses the natural selection conceptualization is *explanatorily more informative.* This is because it adequately refers to those factors that make the difference to the direction of the change on the population level. Furthermore, this description refers to explanatory *types* of properties, thus being more generalizable and giving counterfactual information. It does not matter that the more *detailed* and *accurate* description of the causal processes of evolution is on a different level (the concrete interactions between organisms and their environments). It also does not matter that we cannot identify a difference between those processes that are in accordance with natural selection and those that are not on this more detailed level of description. The description that uses the language of natural selection refers only to a mechanistic *structure*, not causal *processes*, but it still refers to a set of properties and the *outcomes* of these processes. These processes (interactions between the organisms and their environments, reproduction, and the evolutionary history constituted by these processes) are, however, presupposed for there to be a link between the selective pressures and the outcome – they are mechanistic details, often black-boxed. Both causal descriptions (detailed processes and generalizable structures) are correct, but there is a trade-off between being *precise* (regarding what factors are explanatory) and being *accurate* (in the details of the description of the process) in any concrete evolutionary explanation. In this case, however, increasing accuracy decreases *adequacy* as well: we are not interested in how the processes took place, but why they were likely to take the trajectories they took.

Given the framework assumed here, it is sufficient for the logic of the explanation that the relevant causal aspects of the net of ongoing processes are captured. We are not interested in all the causal interactions taking place, but in the reliable net effects on the population level and the properties that have the right biasing effect. Abstracting away from detail enables us to focus on these properties, given that the connection is robust enough. A more detailed description of the causal processes actually *loses* explanatory information. Details of any particular interaction do not tell us anything about why the

population level effect occurs and what is important in those interactions. For example, the advocates of the statistical approach conclude that there is no explanatory difference between natural selection and drift, but this conclusion should be taken as evidence against that approach. There is an obvious explanatory difference: natural selection specifies the conditions under which the probabilistic process of drift become biased.[42] These conditions (abstracted in the form of natural selection explanation) *cause* the bias and explain the evolutionary outcome. For example, if a trait was selected for its use in the environment (it is an adaptation), then if something in the environment changed (*an intervention*) so that the abstraction into conditions of natural selection cannot be done in the new context (for example, the differences in traits are no longer reflected in the differences in fitness), the drift would not be constrained and the direction of evolution would not be biased – the trait would not be selected. This is an unequivocal case of causal explanation. The conditions C1–3 for natural selection are abstractions that specify the causal structure of the interaction that, instantiated in concrete interactions (and the differences between these interactions), constitute an AFI mechanism.[43]

---

[42] To be precise, the drift is not the *theoretical* null hypothesis of evolution: the Hardy-Weinberg Principle, which states that allele and genotype frequencies in a population remain constant from generation to generation in the absence of other evolutionary influences, is (Sober 1984; Stephens 2010). Evolutionary "forces," including drift, are observed through their effects that are measurable deviations from this equilibrium. In any finite population, however, there are random events that cause genetic drift, so this should probably be treated as the *empirical* null hypothesis. (Brandon 2006; see also Futuyma 1998.) But if this is the case, drift should not be considered a "force" or even a causal factor at all (Earnshaw 2015; Luque 2016). The point here is, however, simply that the *change* in population (that is, evolution) can be random (drift) or biased toward some traits evolving rather than other (natural selection), which is an additional factor in the process that still includes drift as an agent of change (see also Hitchcock & Velasco 2014).

[43] Lane DesAutels (2016) argues, against Skipper & Millstein (2005), that natural selection can be understood as Machamer–Craver–Darden mechanisms,

Furthermore, natural selection is an open mechanism: other causal factors affect the process of evolution too. The conditions that constitute the mechanistic structure of natural selection may be present even if the actual causal process is not influenced by these conditions enough for them to be the difference maker. Even in this case, natural selection would have what I will later call a *minimal explanatory role*, since it is a part of the conditions that need to be considered in counterfactual analyses.[44]

I will not go deeper than this into the debate on natural selection as a causal mechanism, for two reasons. First, this is a foundational issue about the nature of evolution that would require much more thorough debate on the foundational questions about causation and causal explanation that I only referred to in the beginning of this section and that I must leave in the background, and it is not the topic of the dissertation. It is sufficient, for now, to outline the mechanistic framework for analysing some of those mechanistic connections in greater detail later. I hope I have sketched an outline of an adequate argument for using this approach. Second, it is fair to say that, although there is still some confusion over the mechanistic nature of

---

if we distinguish between process vs. product regularity, mechanism-internal vs. mechanism-external sources of irregularity, and abstract vs. concrete regularity. I have characterized CCS mechanisms basically as MCD mechanisms, but these liberalizations of decomposability and isolation of the interactions between parts make DesAutel's characterization, in my estimation, an AFI mechanism. The solution still works and is similar to the one presented here.

[44] Denis Walsh and Andre Ariew have pointed out that the current adaptive value of a trait explains (and predicts) the persistence of the trait in the future, even if it was not explanatory for the past evolution, for the same reason: it affects selection in the future (Walsh & Ariew 1996; Walsh 1996). We do not need to go into how this fits their statistical, non-causalist view of natural selection, if it does at all. But under the interpretation advocated here, we can go even further and say that the mechanism of natural selection is currently in place, and this is why we can explain and predict the direction of evolution, within the limits of various constraints and other causal factors.

natural selection, the rivalling statistical view is a minority view,[45] and the third theoretical alternative, natural selection as a law-like primitive, is probably not endorsed by anyone. What is important for what is ahead is that the mechanism of natural selection is an AFI mechanism with parts that are population-level abstractions of processes and differences that take place at the individual level. The population-level descriptions emphasize what is explanatorily relevant in the population structure, whereas the individual-level descriptions of the same processes would not.

[45] For other responses and counterarguments to statisticalists, see Bouchard & Rosenberg 2004, Stephens 2004, Reisman & Forber 2005, Millstein 2006, Shapiro & Sober 2007, Millstein, Skipper & Dietrich 2009, Otsuka *et al* 2011, Ramsey 2013a, and Otsuka 2016. The defenders of causalism concentrate by and large on fitness as the causal factor that makes natural selection a form of causal explanation, whereas I have identified it as one of three causal parts of the selection mechanism. There is no discrepancy between the views, however.

3. Evolutionary Explanations of Behaviour

Let us zoom closer to the main topic of this dissertation now: from what evolutionary explanations are to what they explain about behavioural traits more specifically, and how this is connected to other biological (including psychological) explanations of the same traits. Natural selection explanations have well known constraints both in applicability as a purely evolutionary explanation and as an instrument for knowledge production, especially on human behaviour. Constraints do not, however, imply non-existence, and the topic at hand is on the methodology of such explanations when they are adequate.[46] I will start by discussing adaptationism and the criticism of it, as well as the various concepts of function. The discussion on functions has been about how to identify a function of a trait, which also partly deals with adaptivity and natural selection's role in understanding biological characteristics. The two discussions are connected, although the direction of analysis is different.

I will argue that there are two different issues involved in adaptation explanations. The first is the trait's causal role function in the overall design of the organism in its ecological context from an evolutionary point of view: what does the trait do for the organism's overall fitness in its environment in its form of life, taking its other characteristics into account? This involves an overall (pseudo-purposive) design perspective on the organism, which is justified and explained by the fact that the organism as a whole is a product of an adaptive evolutionary process that has crafted and tinkered it into a somewhat functional whole. This question asks what the trait is *for* in the organism's way of being. The other part is the actual evolutionary history of the specific trait. The latter is a historical explanation, often presented as a narrative that highlights the relevant (but not all) causal factors that

[46] "Evolutionary explanation" is a much wider category than "natural selection explanation", although the first expression is often used when the latter is meant. This unfortunate convention is followed even in the title of this thesis.

directed and constrained the evolution of the trait, including natural selection and the population-level dynamics from selection pressures. This question is about how and for what the trait took its shape historically, from a given explanatory point of view. If these two forms of evolutionary explanations (*evolutionary functional analysis* and *causal history*) fall into one (that is, the evolutionary history is causally directed by the adaptive functionality that we perceive, and the adaptive causal history explains the trait's existence), we have a *true adaptationist explanation* for the trait. This is sometimes assumed to be the case as a working hypothesis for evolutionary explanations of both kinds.

There are good, much debated reasons to think that this presupposition (which was the original target of Stephen Gould's and Richard Lewontin's (1979) attack on "adaptationism") does not work as a universal explanatory strategy in evolutionary history. This in turn casts some doubt on the sensibility of the adaptive functionality analysis part of the question, since the one is justified by the other. If evolutionary functional analysis is to be taken as a causal explanation, a correct form of evolutionary history is presupposed, even if no details of the actual causal history are. In fact, the debate over whether adaptationism is a usable approach anywhere seems to presuppose that these two issues are inseparable. In what follows shortly, I will separate these two questions and discuss both forms of evolutionary explanation, the causal history and the evolutionary functional analysis in the non-historical sense, as well as their relation to other forms of biological explanation (proximate and developmental), against the background of the criticism of adaptationism.

However, when the object of evolutionary explanation is a behavioural trait, further complications emerge. Although the principles of evolutionary explanation apply to behavioural traits just as well as to physical characteristics, behavioural traits are in a more complex relation to the environment from proximate, developmental, and evolutionary points of view. Furthermore, the relation of proximate, developmental, and evolutionary explanations to each other becomes more complex as well. Regarding human social behaviour, an important

factor for this complication is the influence of the social and cultural environment. This is true of each of these dimensions separately, but it is also possible that the interaction between proximate, developmental, and evolutionary dimensions of biological explanation takes place on this level as well, not only through genetic inheritance of behavioural tendencies. To give adequate answers to the questions emerging from this complex, the questions themselves must be framed properly. To do this, I will explicate the relationship between the different kinds of biological causal explanations within the framework that I will articulate shortly. Before going there, I will distinguish between three kinds of explanations involving an adaptationistic approach (or evolutionary functionalism): *evolutionary historical explanations*, *artifact-model functionalism*, and the analysis of the *adaptivity of current use*.

3.1. Adaptationism and Its Criticism

For Charles Darwin (1859), there were two major features in nature that needed an explanation that the evolutionary theory could provide: the family-like similarities between species (or *unity of type*), and the fact that biological beings seem to be designed for their environment. The first question is answered by common ancestry and the stability of certain developmental and structural features from generation to generation (*phylogenetic inertia*), and the second one by the mechanism of natural selection. (See Depew & Weber 1996; Gould 2002.) Evolutionary explanations are historical explanations for why certain traits exist in the population and why they possess certain general characteristics. Natural selection explanations are evolutionary explanations that refer to natural selection as the primary explanatory factor, in contrast to, for example, drift, or *developmental constraints*.[47]

[47] Developmental constraints are the factors in the individual development that limit evolutionary change either by limiting the range of phenotypic variants that are possible in the ontogeny (*generative constraints*) or by limiting the viability or

The developmental constraints both conserve the developmental forms from earlier evolutionary history (resulting in *homological* similarities and partly explaining phylogenetic inertia) and bias the direction evolutionary change by constraining the production of variation for selection to work on (see Maynard Smith et al 1985; Amundson 1994 & 2001; Richardson & Chipman 2003; Pearce 2011). There are also random factors that affect evolution, such as drift and mutation. Since they are not directed, their effect gets smaller over time, but they do have consequences, especially on small populations (Futuyma 1998). Moreover, adaptation takes time and is path dependent. Natural selection explanations are not the only evolutionary explanations. In what follows, I will give a very brief review of some of the critical discussion on adaptationism. My intention is not to participate in this discussion as such. The clarification of various possible adaptationist stances and their criticism is a starting point for developing my own account of evolutionary functionalism in explanation and helps to clarify what I am proposing.

3.1.1. Kinds of Adaptation Explanations

In the previous chapter, I defined natural selection, following Endler (1986), as a process in which:

C1) there is *variation* in the given phenotypic trait on the population level,

C2) there is a consistent relationship between the trait and *fitness*, and

C3) the trait is *inherited*, and because of this,

O) there will be a change in the frequency of distribution of the variation in the trait between generations.

---

survivability of the organism during ontogeny (including through the consequences for other developing traits; *selective constraints*) (Richardson & Chipman 2003).

C1–C3 constitute an explanatory mechanism for O. Note, however, that the condition C2 only states that there is a consistent relationship between the trait and *fitness*, or, that the expected reproduction rate relative to others in the population. It does not specify *why* the relationship holds. To establish this, we need to specify a mechanism linking the trait and fitness. A causal (non-arbitrary) connection could hold because the trait

C2a) is *adapted* to its environment (that is, it has benefits for the survival of the individual possessing it in that specific environment),

C2b) enhances the individual's reproductive capacity (fertility, mating success, and so on),

C2c) has a supportive function for another trait that fulfils C2a or C2b directly, or

C2d) is otherwise associated with another such trait (for example, because of a structural or developmental connection such as pleiotropy).

C2a–c are cases of the trait itself causally contributing to the fitness and therefore its own existence. They are selected *for* instead of just being selected, as in C2d (see Sober 1984a). Being adapted (C2a) is a qualitative ecological notion that refers to the fit between the organism or its trait and its environment (Brandon 1978). An organism can be more or less adapted (and in many cases variation can be quantified, and one can make optimality models for measurement), but the notion is about the qualitative fit, having a function, in contrast to the quantitative notion of fitness. Fitness is a measure for reproduction capacity of an entity that reproduces: depending on what the theory in use allows, a gene, an individual, or a group. The fitness of an individual (or a gene) supervenes on the individual's traits (or the traits associated with the gene) depending on their overall adaptedness (C2s) as well as the factors in C2b and C2c. Fitness is not a causal

property itself: it is a measurement of relative differences between reproductive units in the causally relevant properties that direct evolution by natural selection. (Futuyma 1998; Sober 1984a: 97–102). What is usually meant by "adaptation" is either the process in which a trait evolves because it gets selected for its adaptedness, or the trait produced by this process. The adaptation explanations rely not only on the mechanism of natural selection (which covers all C2s) but also presuppose that the trait has something it has been selected *for*, its *adaptive function*. In the strictest interpretation only C2a qualifies as an adaptation since it is the only one in which the adaptive value of the trait is causally related to its origins. In a more inclusive reading, explanations referring to C2b or C2c can also be included in the adaptation explanation, although they may have originated for some other adaptive purpose or for none at all, and have only acquired a new role. Stephen Jay Gould and Elisabeth Vrba (1982), for example, distinguish true adaptations from such traits, which they call *exaptations*. I will argue later why a more inclusive reading is more sensible. Traits that are selected because of an association (C2d) but have no selective function are by-products.

Not all evolutionary explanations are adaptation explanations: some are by-product explanations, some refer to drift or phylogenetic inertia, and some may refer to historical contingencies in the evolutionary pathways. "Evolutionary explanation" is a larger category, "adaptation explanation" is a sub-category. Adaptations and natural selection have, however, a central place in evolutionary explanations. In what follows next, I will highlight the reasons why and under what limitations. I will also discuss the explanatory relevance of adaptive functionality extracted from a mere causal-historical explanation. Evolutionary adaptation explanations are interested in history, but a snapshot of the current adaptation process may be interesting in some cases, too. Moreover, as I mentioned in my discussion on mechanisms, an explanatory causal structure (such as the factors constituting a natural selection) may exist without a causal process going through it. More importantly, there is another reason besides the explanatory

relevance for why we think that C2a is special among the C2s: it is related to how the organism fits its environment. I will argue that this is in itself an important perspective for understanding living beings, not just because of its role in evolutionary-historical explanations.

3.1.2. The Problems of Adaptationism

The methodology of studying practically all biological traits as adaptations, or presupposing they are adaptations, was labelled "adaptationism" by Stephen Jay Gould and Richard Lewontin (1979) who began to criticize it as a flawed practice (see also Gould & Vrba 1982; Gould 2002; Lewens 2009). Adaptationism is, however, several things wrapped into one package. I will now distinguish different concepts of adaptationism. This is not done to sort out different *views* – in fact, the "adaptationisms" discussed next are not different stances or views about adaptation, and they are not theoretically independent either. The aim is, instead, to sort out the overall baggage of adaptationism that is a net of connected ideas about the nature of biological evolution, explanatory choices, and methodological tools. The purpose of this is to find out how evolutionary explanations could and should be understood in the context of social behaviour and what the consequences of this are for the issue of individualism vs holism in the human context. A starting point for distinguishing between various ideas related to it is Peter Godfrey-Smith's (2001) distinction between *empirical*, *explanatory*, and *methodological* adaptationism, that can be further refined into several alternative interpretations of the subject matter; I will mostly follow Tim Lewens (2009) in this. Furthermore, a fourth important aspect of adaptationism (that was also discussed in Gould & Lewontin 1979) is the issue of *atomism and holism*[48] about the characteristics of an individual: how isolated from the overall

[48] "Holism" as in the holistic nature of an individual organism, in contrast to an individual being a mosaic of traits that perform their evolutionary functions in isolation from each other.

architecture of the organism and its behaviour can we presuppose the traits under explanation to be in their adaptive function?

Some of these forms of adaptationism or distinct contents of adaptation claims can be distinguished into further sub-categories. The forms of adaptationism that I will discuss are the following, with some preliminary characterizations:

Empirical adaptationism	**-** *all traits are products of adaptation*[49]
Pan-selectionism	*- natural selection overrides other forces*
Good-designism	*- the overall design of organisms is a good fit to the environment*
Well-enough-designism	*- the overall design is functional in the environment*
Gradualism	*- the evolution of a trait is about a gradual change*
Explanatory adaptationism	*- the replies to evolutionary questions include a reference to natural selection*
Strong historical expl. adaptationism	*- in evolutionary history, explanatory questions will have an adaptation answer*
Weak historical expl. adaptationism	*- in evolutionary history, explanatory questions will have an answer that* ***includes*** *natural selection*
Ahistorical expl. adaptationism	*- adaptive functions are explanatory for the functioning of the organism*

[49] This may be more or less inclusive in what counts as "adaptation", as discussed above. The definition has implications for what adaptationism in this sense means, but we can disregard this for the current discussion.

Methodological adaptationism	*- all traits should be treated **as if** they were adaptations*
Historical meth. adaptationism	*- traits should be treated as if they were historically selected for what they do*
Ahistorical meth. adaptationism	*- instrumentalistic adaptationism in the functional description of the organism*
Atomism	*- traits are treated in isolation of each other regarding adaptivity*

Empirical adaptationism is a claim about nature and its history, about the causal factors guiding the evolutionary history. According to it, everything (or almost everything) in nature serves an adaptive purpose and has been selected for that. In other words, it is an empirical claim about the relative powers of natural selection and other factors in evolution, and it is connected to actual historical processes. Whether or not this is so is a part of the research object of evolutionary biology proper, and from that perspective, this is the main question. Furthermore, it is the idea that most criticism of adaptationism has been targeted against.

There are two sets of possible alternative factors for natural selection to be actual historical reasons for a trait to exist (see Gould & Lewontin 1979; Futuyma 1998). The first set consists of chance events. Since both populations and the time that they have to adapt to a particular environment are finite, the right mutations may or may not appear. Furthermore, there is a chance element in the order of mutations appearing, which may be consequential, and natural selection is path-dependent in any case. Drift and migration may have consequences depending on the relative strengths of these factors and the selection. The other set of factors comes from the holistic nature of organisms and their development that constrains natural selection (see also Amundson 1994 & 2001; Gould 2002). There are structural dependencies between functionally different parts of an organism, the

developmental processes may connect both functionally and structurally separate traits, and some path-dependencies in the development make some changes in the organism's design impossible. Empirical adaptationism is in its essence an empirical thesis about the strength of natural selection in relation to these factors. This in turn contains a whole set of presuppositions. Tim Lewens (2009), for example, further distinguishes three conceptually distinct presuppositions within empirical adaptationism: *pan-selectionism*, *good-designism*, and *gradualism*.

Pan-selectionism is the empirical adaptationist claim about the strength of natural selection in relation to the chance factors: natural selection is the strongest causal factor acting on the population level. The development does not matter, since this is a thesis about what happens to the variation that is available in the first place.[50] A complementary empirically adaptationist claim would be precisely about the strength of natural selection over developmental factors. This claim would say that natural selection can break the links between functionally different traits and overcome all possible developmental constraints. This position is probably embraced by no one, since it is simply an absurd claim from a biological point of view. But there is a weaker version of it, which Lewens calls **good-designism**. According to this idea, organisms are *overall* well adapted to their environment, even if many of their characteristics are not produced by natural selection for the purpose to which they are being put. I will return to this idea later in more detail. Pan-selectionism and good-designism do not entail each other. The third form of empirical adaptationism that Lewens discusses is **gradualism**, a claim that all apparent design in the nature is produced by gradual evolution guided by natural selection for the particular purpose that they are designed for. This rules out things like

[50] Developmental constraints (see Amundson 1994 & 2001) are, of course, constraining possibilities for the evolutionary process, but natural selection does not produce variation but rather selects from it, so this should not be thought to constrain natural selection as such (Orzack & Sober 1994a, 1994b & 1996; Sober 1998b). Developmental constraints are constraints for *evolution*, but not for natural selection.

an evolved trait adopting a new function (or, to use a term proposed by Gould & Vrba 1982, becoming an "exaptation"). Gradualism is independent of the other two forms of empirical adaptationism.

The problem with all three claims of empirical adaptationism is that they are empirically false.[51] There are other population-level factors than natural selection, organisms are not always optimal for their environments or in their design – biology is full of bad design rooted in structural and developmental constraints[52] – and traits often adopt new functions (see Futuyma 1998). But it is important to keep these complications separate and distinct even in non-adaptationist methodological considerations about the role of adaptation explanations and adaptive functionality, and I will return to them in the next section. At the same time, even many of the most ardent critics of adaptationism think that natural selection is, in some sense, the most important factor in evolution (see Gould & Lewontin 1979; Gould & Vrba 1982; Gould 2002). The point is not that adaptationist explanations are invalid or unimportant, or that true adaptations would be rare, or anything like this. The point is that there are other factors that should be taken into account, the developmental factors being the most important. One of the main targets of criticisms in Gould and Lewontin's original anti-adaptationist paper was that sometimes some

[51] See Futuyma 2010 for a review of empirical adaptation "failures" and some of the known and speculated reasons for this.

[52] For example, there are vestigial parts and structures that have lost their usage and have become burdensome. An often-used example of this, because of its striking dysfunctionality, is the anatomy of giraffe's neck. First, there are developmental constraints that limit the number of neck vertebrae that seem to be specific for mammals but not, for example, for birds. Almost all mammals (with the exception of sloths) have only seven cervical vertebrae, which makes their necks unnecessarily vulnerable from a purely design point of view. Second, their recurrent laryngeal nerve makes a long, unnecessary loop that is dysfunctional. (Badlangana, Adams & Manger 2009.) Traits like this call for an evolutionary explanation that describes the non-adaptive contingencies of the actual evolutionary history.

evolutionary biologists simply presuppose the traits under study to be adaptations, and if a hypothesis fails, another adapatationist hypothesis is made, while the correct explanation for the trait's existence is not an adaptationist one in nearly all cases.[53] This cannot be accepted, of course: even if a trait being an adaptation for something is a primary hypothesis in studying its evolution, the fact of it really being so should be a result of evaluating evidence, not a presupposition guiding the interpretation of empirical data. This is simply lazy and methodologically bad science.

3.1.3. Adaptationism as an Explanatory Perspective

Abandoning crude empirical adaptationism, there are still other possible justifications for granting adaptation and adaptive functionality a special role in research, from an explanatory or methodological point of view. Some proponents of adaptationistic approaches highlight the uniqueness of natural selection as an explanatory resource, because it is the only thing that can account for *adaptive complexity* (Dawkins 1983 & 1986; Dennett 1995; Sober 1998). Traits could be adaptive by chance, and evolving systems tend to become more complex over time, but the only known factor that guides the evolving complexity systematically towards adaptive organization is natural selection. According to some, this gives adaptationist explanations a kind of priority: natural selection is not the only causally contributing factor, but it is the difference-maker for the questions in which we are interested most in evolutionary biology. This is the view Godfrey-Smith (2001) calls

[53] In practice, if there are several possible evolutionary hypotheses, only adaptationist ones are generated or taken seriously, or adaptivity itself is considered evidence for the hypothesis. Elisabeth Lloyd (2006) provides a revealing case study on this practice. She points out that the existing evidence seems to back a by-product hypothesis on the evolution of human female orgasm, while the various adaptationist hypotheses are almost exclusively discussed in the literature.

explanatory adaptationism, and as he remarks, it is more charitable to consider many adaptationists explanatory rather than empirical adaptationists (and this would probably include Richard Dawkins 1983 & 1986 and Daniel Dennett 1995). I will later make a further distinction between *historical* and *ahistorical* explanatory adaptationism.

As mentioned above, much of the criticism from Gould and his collaborators is aimed at the presupposition that, from a causal-historical point of view, *some* adaptationist explanation is going to be correct. This strong version of historical explanatory adaptationism would be the thesis that all adaptedness to the environment is in fact selected for and it would presuppose (all forms of) empirical adaptationism. A weaker version would be a view that acknowledges other factors but gives adaptation a special status: even though there are multiple causal factors in the evolutionary past, natural selection has an explanatory priority. As Godfrey-Smith (2001) and Lewens (2009) point out, the formulations that fall under this view (for example, Dawkins 1986) often sound like statements of what is *interesting* (versus boring) as an explanatory question, or what is the "proper" object of study within the discipline. This makes the question something of a matter of taste and an unnecessary normative constraint on research and discovery. In a more charitable reading (of, for example, Dobzhansky 1973 and Dennett 1995), this idea can also be taken as a philosophical point about understanding how adaptive complexity or apparent design is naturalistically possible in the first place (see also Dawkins 1983). This, in turn, is undeniably true, but not relevant to any specific historical explanations as such. Another interpretation is a substantial thesis about *explanatory relevance*. This reading is not trivially wrong in the light of the falsity of empirical adaptationism alone. This is also where the difference between chance factors in evolution and the holistic nature of organisms as arguments against adaptationism discussed above become relevant. For even a weak form of historical explanatory adaptationism to be true, it has to overcome (even if not debunk) the objections to the empirical adaptationism, and the two different groups of objections discussed above call for different answers.

Pan-selectionism is not true, but it is not entirely false to say that much of what is in nature gets selected for a "purpose". Even in the presence of strong selection, drift and other chance factors have some effect, but they do not have a uniform direction. Natural selection has a qualitatively different explanatory relevance in understanding long-term trends in evolution: it gives evolution its direction. Although drift is sometimes the difference-maker between two possible evolutionary paths, this is more likely when there are no strong selective pressures, and the population is small. Furthermore, natural selection plays some role in any complex adaptive system, and it is the factor behind the purposive-looking design – even if it is not the only factor and the apparent purposefulness was not the selected function. In addition, building an optimality model based on what a trait would be if completely adapted is the best way to find out about the existence of chance (and other) factors. Furthermore, we can understand the mechanics of adaptation process and build generalizable, testable models about them – it makes sense to start from there. (See Brandon 1978 & 1990; Stephens & Krebbs 1986; Orzack & Sober 1994a, 1994b, 1996 & 2001b; Sober 1998b.) For these reasons, it seems that the critiques of empirical pan-selectionism are not crucial challenges to granting adaptationist explanations a special explanatory and theoretical role, as long as other factors are taken into account and the empirical reality is such that adaptation processes are commonplace. (They are.) One could view the evolutionary shaping of a trait as a dialectic between natural selection and chance, and from a theoretical point of view, natural selection plays a special role. This may be called **weak historical adaptationisim** in contrast to the stronger idea that natural selection is the only explanatorily relevant (or interesting) factor.

There are many causal factors in any historical event, evolutionary or otherwise, that need to be described for a full understanding of what exactly happened. This is the case even if we are only interested in *relevant* information, not in *complete* description. Some of them, however, are more relevant than others, depending on the explanatory interests and the chosen framework (under the view on explanation assumed here).

The events and causal processes remain the same, but which ones are important to us (or explanatory of things that are important to us) will change. If we do not give the full story, different explanatory perspectives may highlight different aspects of the ancestral environment to explain the same trait without being in competition. One should be just as pluralistic about evolutionary explanations as about any historical explanation, and for similar reasons.[54] If we are interested in why birds have wings, we may very well be interested in several different causal factors in the evolution of bird wings, and different research interests imply different factors to be important. We may concentrate only on evolutionary adaptive functions or on all causal factors, and even within the adaptationist framework, we might only be interested in those factors that are related to flying even if other selective pressures were crucial in the beginning of wing evolution. There is nothing wrong as such in constraining the explanatory interest in this way. This becomes a problem if other factors are denied (the traditional accusation toward empirical adaptationism) or the chosen perspective requires idealizations or other simplifications that affect the factuality of the case.

---

[54] This is a familiar issue regarding history proper. For example, if we are interested in understanding the fall of some European countries under authoritarian rule in the 1930s, there are numerous historical facts that played causal roles, but only some seem to be more relevant to making sense of the *difference* between *which* countries fell under authoritarian rule and which remained (or developed into) democracies (see Mann 2004). We can isolate these factors as explanatory even if they give only fragments of the actual causal history. Furthermore, in any such historical event, we are often interested in the difference-makers that explain those aspects of the event that are *important to us* from our contemporary perspective. For example, we may be interested in how the societies falling under authoritarian rule differed from our contemporary European societies (some of which seem to be surprisingly fragile democracies) instead of comparing them to the democracies of their time. This is the "teaching" function of history, and why every generation *needs* to write its own history, as it were. The events and causal processes remain the same, but which ones are important to us (or explanatory of things that are important to us) will change. This can also be extended to *evolutionary* histories.

A more crucial challenge that is lethal for even a weak historical explanatory adaptationism comes from the holistic nature of organisms and their developmental constraints (see Amundson 1994 & 2001; Gould 2002; Pearce 2011). This has consequences for how we should understand natural selection even if there were no competing causal factors affecting the change on the population level. Limitations in the possible variations, linkages between different traits in development, and the practical impossibility of some changes due to dependencies between the traits or their development conserve "received" forms in, for example, body plans. Changes like mammals getting an extra pair of limbs to function as wings, or whales evolving gills, are nearly theoretical impossibilities. This fact, which is sometimes called *phylogenetic inertia* (Harvey & Pagel 1991; Orzack & Sober 2001b), performs much of the explanatory work in understanding living beings. Even if adaptation were the central explanatory interest as a starting point, evolutionary biology has to account for when adaptationism does not work and when there are systematic "failures." Evolutionary biology must be about these factors, too (see Futuyma 1998). However, the inadequacy of adaptationism goes deeper than this. First, as mentioned above, evolutionary theory has two general explanatory interests in its task of explaining why biological beings are the way they are: not only their apparent design, but also the family-like similarities between species. Phylogenetic inertia and developmental constraints are the explanatory basis for the latter. Mere common ancestry is not enough: it provides the starting point, but the developmental constraints are needed to explain the extent of conservatism in evolution. Second, phylogenetic inertia plays a substantial role in the historical explanation of traits even if the focus is on adaptation. This is because in the study of adaptive evolution, the optimality models need to be combined with comparative methods, which include both analogical and homological comparisons (Orzack & Sober 2001b) – that is, both comparisons between independent but similar evolutionary paths on different lineages and comparisons between the "shared" traits of two closely related species to find

evidence for evolution. However, this topic as such is not important for the dissertation at hand.[55]

A possible reply to this problem is to be flexible with the timeline and concentrate on the *original* adaptivity of the phylogenetically inert traits (Reeve & Sherman 1993 & 2001). Since even phylogenetically inert traits (may) have been selected for some adaptive function in the evolutionary past, the reply goes, this is the relevant fact we need to know. If we want to understand why a trait exists, it is not enough simply to say that it has existed before and now it is stuck in the body plan; we need to understand its origins. This moves the adaptationist question back in evolutionary history and away from current use and gives the adaptationist perspective on an organism historical depth. This is also the adaptationist response that evolutionary psychologists gave to some adaptationist problems of human sociobiology. However, as a general solution, this fails. One could be interested in the origins of a given trait existing across a clade, for example, but if one's historical adaptationist explanatory interests are in explaining traits of a particular population (for example, humans), given a historical starting-point (for example, after the separation between our own line and the chimpanzee's) and the selection pressures from the evolutionary environment, then the explanation is the trait's pre-existence and fixity in development, not the previous history.[56] Furthermore, and most

[55] If two species share some characteristic because of their common ancestry, it is a *homology*. There has been debate on the nature, definitions, and relevance of homologies in the philosophy of biology. Some even think that "homology thinking" (Ereshefsky 2007) is just as important as the design perspective (see Griffiths 2006; Ereshefsky 2012). I agree, but this issue is not of relevant for the topic of this thesis.

[56] Whether or not this works with evolutionary psychology is a bit more complicated. The point of much of evolutionary psychology is that there is an environment which is the proper environment of interest for evolutionary questions (the "environment of evolutionary adaptedness"), and all adaptation that is important for understanding human psychology took place there, not in contemporary environments. This does not confuse adaptation with

crucially, even if a trait's mutability is constrained, selection may still have tinkered with it substantially, so that the end result is not perfectly adapted to anything in particular – not to the current environment, and not to the ancestral environment of the clade's origin either. This sub-optimality (for everything) requires an explanation that refers to developmental constraints. An evolutionary historical explanation needs to integrate adaptation and phylogenetic inertia.

Another line of argument is that we should consider the developmental environment of a trait an internal environment that the trait has to be adapted to (Reeve & Sherman 1993; 2001; see also Wimsatt & Shank 1988). In this way, explanations with developmental constraints would be just a type of adaptationist explanation. Nevertheless, this will not work either. As Lewens (2009) argues, this presupposes that alternatives that get selected against are developmentally possible, which is not always the case either (Amundson 1994 & 2001). And even if it were the case, this would already expand the perspective from mere external function of the trait to the developmental factors, which is exactly the point of the criticism of adaptationism in the first place. In other words, developmental factors are simply too central to be dismissed.[57]

3.1.4. Adaptationism as a Methodological Tool

Explanatory adaptationism therefore does not seem to be a viable position either. This does not mean that adaptation explanations are not important in historical explanations – they are, but the adaptation process is only one causal factor and should be treated as such. There is, however, a third way to see the primacy of adptationist explanations, as a methodological tool. Godfrey-Smith calls this **methodological adaptationism**. Because natural selection is a central factor and tends to

previous adaptation history as such, but is a similar general move in saving the adaptivity by moving back the context of adptivity, nevertheless.

[57] I will come back to importance of development in chapter 7.

figure in the explanations of apparent design in nature, some think that the search of explanation should begin from there, as I have already pointed out. Looking for an adaptationist explanation through optimality models and selection scenarios may be an important method to discover developmental constraints, too, through failures in the predictions (Brandon 1978 & 1990; Mayr 1983; Orzack & Sober 1996 & 2010a). Elisabeth Lloyd (2015) has, however, argued convincingly that this methodology falls back to the other forms of adaptationism: since the central question is still *what* the function of a given trait is, not *whether* it has a function, only adaptationist explanatory hypotheses are tested and the evidence becomes biased (see also Green 2014 for similar worries).

I will argue in the next sub-chapter, nevertheless, that it is not necessary to interpret all adaptationist explanations as purely historical explanations of the original evolution of the traits. An alternative is to see the adaptivity of a trait as its function in the whole of the organism in relation to its environment in evolutionary terms: an adaptation is a state of an organism, independent of historical origin, which might or might not be a historical adaptation as well (see for example Fisher 1985). The rationale for this is as follows.

If organisms tend to be adaptively functional in their environment, it could be useful to approach their functional organization from this point of view, *as if* the organism were adapted into its environment, even though it was not adapted into it historically. The tendency itself may have many different historical explanations. Besides adaptation and exaptation, developmental plasticity, and environmental induction (West-Eberhard 2003), as well as evolution through non-genetic inheritance mechanisms (culture, behaviour, epigenetic inheritance; Jablonka & Lamb 2005), increase adaptivity. The causes for adaptivity are not in one direction, either: organisms can modify the environment they are adapting to (*niche construction*; Odling-Smee et al 2003) and organisms may select the environments in which they live instead of the environments selecting the characteristics of the organisms (*habitat selection*; Morris 2011). All these processes can increase the

adaptedness of organisms to their environment in a dynamical relation with the environment, without being cases of strictly Darwinian selection processes.[58] This idea, too, should fall under methodological adaptationism in Godfrey-Smith's (2001) taxonomy. I also take something like this position to be the methodological stance advocated by Daniel C. Dennett (1995). For Dennett, the core of adaptationist heuristics is that they are useful, and they provide justification for the "design stance" (Dennett 1987), an approach to living organisms as if they were designed, which is necessary for us to make sense of them.[59]

As Lewens (2009) points out, the first point (the usefulness) is an empirical prediction that can be roughly equated with what he calls well-adaptedness. Lewens bypasses the second point (design perspective) as mere "heuristics", but here I disagree. It may be viewed as a stronger point that is related to what I have already mentioned, calling it *ahistorical explanatory adaptationism*. The point is that the adaptive function of a trait has an explanatory role even if it is not the actual historical cause for it, as a form of functional explanation.[60] There is, of course, the semantic issue of whether this should still be called

[58] The so-called "extended synthesis" (Pigliucci 2007) seeks to integrate developmental and ecological aspects into evolutionary theory and form a new paradigm for studying evolution (see also Pigliucci & Müller (eds.) 2010; Laland et al 2014; Laland et al 2015). Rather than being a challenge to adaptationism, this new paradigm could be interpreted as an extension of what explains the adaptation of organisms to their environment. I will discuss the extended synthesis later in relation to evolution and development, but the ecology aspects are an equally important extension.

[59] This idea of design without a designer, or purposiveness without purpose, towards living nature as a *necessary* tool for understanding it, given how human mind works, goes back to Immanuel Kant (1790), interestingly enough. (See Depew & Weber 1996; Grene & Depew 2004.)

[60] Lewens (2004), Forber (2009) and Green (2014) have also pointed out that a weaker form of adaptive thinking, as a heuristic to discover the capacities of an organism and their functionality, does not require a connection to historical adaptation and should be treated separately. I will argue in what follows that these are two quite different "evolutionary" approaches that should not be confused.

adaptationism, even if the view is defendable. This is not important, but to avoid confusion, I will call this combination of certain aspects of methodological and explanatory adaptationism *evolutionary functionalism*. In what follows next, I will explicate this idea as a positive position on the role of adaptation explanations in non-evolutionary biology, including behavioural biology, and I will contrast it with other forms of adaptationism. In practical terms, I divorce evolutionary functionality and historical causes, which may come together (in true adaptationist explanations), but they may not. I will start by discussing the various concepts of biological function, many of which are to specify what the given trait is *for*, with or without the idea of the trait having been actually selected for that task, and I will discuss how adaptivity figures in this debate. After this, I will distinguish three different forms of evolutionary functionalism: *current use*, *historical evolutionary functionalism*, and *ahistorical evolutionary functionalism*.

3.2. Evolutionary Functionalism

Much of biological explanation is functionalist. Traits and processes are understood by their purpose for and within the organism, and this purpose is naturalized in one way or another. One such way is evolutionary functionality. This can also be a perspective on social behaviour, including human behaviour. If this is the chosen perspective, questions arise. Functionality of what? For what? Functionality in what sense? Some of these questions will lead to the questions about the relation between individual and above-individual level descriptions and explanations.

3.2.1. Adaptivity and the "Consensus without Unity"

One way to characterize evolutionary functionality – or even to give a general explication for the biological notion of function – is to say that *a trait can be explained as a component in a mechanism that produces*

something that is useful for the organism. This participatory role in a mechanism is the trait's function. There are different ways to understand what exactly this means, and there is not even a clear view of what the analysis of function is supposed to achieve. Functions seem to be useful in distinguishing between purposeful and accidental (for example, the heart pumping blood versus making noise), they are supposed to be explanatory (why we have a heart), and they have a normative aspect, that is, distinguishing dysfunctional (not pumping properly) from functional, and malfunctioning (ventricular fibrillation) from functioning, and all these dimensions of functionality are contested (Garson 2016). Pluralism about function concepts is widely accepted (see Godfrey-Smith 1993; Amundson & Lauder 1994; Griffiths 2006; Bouchard 2013; Garson 2016). As Peter Godfrey-Smith (1993) put it, there is a "consensus without unity." I will follow this consensus and, rather than defending a view of functions, I will present the "basic concepts" of function and discuss them in connection with our current purposes. The three concepts I consider to be the basic concepts of function are the *etiological function***,** the *causal-role function,* and the *adaptive function*.

The **etiological function** of a trait (Wright 1973 & 1976; Millikan 1984; Neander 1991) is whatever the trait does that explains its existence. Larry Wright (1973: 161) defines it thus:

> "The function of X is Z means: a) X is there because it does Z and b) Z is a consequence (or result) of X's being there."

The reason for including *X doing Z* as the cause for the existence of X is to naturalize the apparent purposefulness and explain it as well as to provide criteria for distinguishing the "real" function from everything else that results from X. There is no theoretical necessity to restrict the use of the concept of etiological function to the evolutionary dimension of biological causes, but in practice, and especially within the evolutionary perspective, a biological trait has etiological function only if it is an adaptation. Ruth Millikan (1984) and Karen Neander

(1991), for example, defend the **selected effect** variant of etiological function. In this view, the function is whatever it does so that the organism becomes fitter, which in turn has caused the trait to be selected. Whatever else it does for the good of the organism is just a set of other consequences. This is also a way to understand the adaptationism debate. Empirical adapationism is a claim that (most) traits have an etiological function, and both explanatory and methodological adaptationism give a trait's etiology a special role in describing the trait. If this were the only way to understand functions, many biological traits, including some vital ones, would not have functions at all: useful by-products and exaptations could not be considered to have functions.

The **causal-role function** (Cummins 1975 & 1983) is an articulation for a function that is used to analyse the trait from the point of view of its causal role in the system that it is a part of, from the point of view of the functioning of the whole system. It is a more liberal account that does not require the function to contribute to its own existence. The proponents of this approach argue that there is a need for a concept of function like this, especially in physiology, neuroscience, ecology, and ethology (see Amundson & Lauder 1993; Bouchard 2013; Craver 2013).[61] The causal role function of a trait cannot be whatever causal consequences it has as a part of a system. Etiological functions are too demanding for many purposes, but causal role functions require a perspective from which the functions of the parts are analysed to rule out accidents and unimportant effects. Therefore, the definition needs to include a perspective (see Cummins 1975 & 1983):

> An *X* has a function *F* in the system *S* in relation to an analytic perspective *A* that is adequate to the *S's* capacity to *G*
>
> *if and only if*

[61] The different notions of function should not, however, be automatically taken to be differences that follow from the different fields of biology (Garson 2016).

> *X* has a capacity to produce *F* (or have a role of *F*) in the system *S* as a part of a process resulting in *G* under the description within *A* that has a reference to an *X* capable of *F*.

Unlike with etiological functions, there is no objective way to say how to choose *A* and *G* given any *S* and *X*. An obvious perspective is whatever explanatory interest we have (Rattcliffe 2000; Lewens 2000), which frames the contrastive-counterfactual questions we ask, and, therefore, which causal functions are constitutive parts of a mechanism we *use* for explanations. Therefore, we can say that:

> An *X* has a function *F if and only if*
>
> 1) there is an effect *G* that
> 2) we are explaining from the perspective *A*
> 3) with a mechanism *M*
> 4) that has *X*
> 5) with a capacity to *F*
> 6) as a constituent part of *M*.

Causal role functions refer to real causal relations, capacities and processes, and are explanatory of the whole system and its capacities (they tell what the part does from the perspective of a wider system), but not all of the functions of a trait are equally interesting from the specific explanatory point of view. For this, we need some further criteria. An important factor is that biological organisms are functional wholes that have architectures, as discussed above. Whatever perspective a researcher takes, this internally functional architecture and its mechanistic constitution and maintenance (that is, how the organism stays alive) are a natural part of that perspective. An important dimension in the meaning of "function" is that the trait contributes to this instead of being dysfunctional (for the whole) or malfunctioning

(that is, causing harm to the whole organism by not operating in an orchestrated manner).[62]

The overall architecture, however, is not only a functional whole in itself. It must be functional in its environment, too – as a whole, even if not all its characteristics are functional. The organism might not be optimal, and it might not be "well-adapted", but it has to have the means to operate and survive in its environment. Even if the organism's design is widely sub-optimal, how it nevertheless survives in its environment is what is important in explaining the organism's characteristics. For example, whales having lungs is not exactly an optimal solution, and the explanation for their existence is in the contingencies of the evolutionary past, but understanding the characteristics of the whale lungs and the whole respiratory system must include adaptationist thinking – albeit acknowledging that the functionality of the system centres around something that is highly dysfunctional itself. Much of the design of a given organism may be due to phylogenetic inertia and the path-dependent nature of evolution, but some of this is adapted and some parts have been utilized for new functions (as exaptations). These characteristics are not (historically) adaptations for their use, but they are, nevertheless, a part of the organism's functional design from the point of view of its overall functionality in its environment. The architecture needs to stay, in the course of evolution, an orchestrated whole, and functional enough in the environment the organism lives in for it to be able to live in it, even if sub-optimally. If the biologists are interested in describing the functional

[62] This is not always the case. If we want to understand the biology of a disease or impairment, it is not enough to know that something is malfunctioning from this point of view (although it is the starting point; see Wakefield 1992; Adriaens & Andreas De Block 2011), but we might be interested in how the disease, a cancer for example, functions from the point of view of what it is doing that is harmful for the organism. Given knowledge about some part of the "mechanism" of cancer (for example, its ability to form an internal blood circulatory system), we can try to prevent it growing based on this knowledge (see for example Tammela *et al* 2008).

architecture of an organism, it is only reasonable to include an ecological perspective, which in turn is closely connected to evolutionary considerations. Furthermore, mechanisms used in biological explanations require an idea of what the mechanism is for, and a functional analysis is needed for this (see Craver 2013; Garson 2013).

Whatever role the trait plays in the overall adaptive functionality of the organism in the environment in which the organism lives is a natural starting point for understanding the organism. Even if the adaptive functionality from this point of view is not the result of an *atomistic (historical) adaptation*, an *ahistorical adaptationist perspective* may be a useful heuristic to decompose the organism's functional architecture in its ecological context. Exaptations and structural characteristics that are phylogenetic relics can be treated as functional properties of the overall design – if one steers clear of making *historical* claims or *optimality assumptions* about them. This also seems to reflect real biological practices: biologists are interested in the adaptivity of an organism and its traits in its current environment as an essential part of understanding the organism without speculating about their evolutionary history (see Amundson & Lauder 1994; Walsh 1996; Wouters 2003; Cuthill 2005; Sherry 2005). Justin Garson (2016), however, voices a suspicion that this practice may implicitly rely on adaptationist thinking. This concern is warranted (see also Cuthill 2005), but regardless of what some biologists may presume of the connection of adaptive functionality and historical adaptation, there are other justified reasons to be interested in adaptive functionality alone.

For starters, there are other reasons than a historical adaptation process that lead into ecological adaptivity, as I mentioned above. First, there is *habitat selection*: some organisms can choose an environment (recognizing environmental clues; this is *habitat choice*), or they may randomly end up living in an environment in which they are ecologically more functional, with increasing reproduction rates. Similar adaptive optimization takes place, only without any change in the organism, and an adaptationist perspective can be a useful ecological perspective even when we *know* that the organism has not evolved in

that specific environment. (Morris 2011.) Furthermore, if there are constraints on adaptation in some environment, another environment may be a better fit for the species: either it is directly better, or it is easier to adapt to within the given evolutionary constraints. A whole species may change its niche or migrate to another region – which is an adaptive process, too, producing adaptive functionality without organism-changing (Darwinian) evolution. This might happen when there is a change in environments (for example, a climate change) and a population changes its geographical location to keep its habitat. Organisms may also find new environments that fit them as well as or even better than the environment they evolved in.[63] Furthermore, generalist organisms with behavioural plasticity may be able to utilize the features of the environment that are utilizable for them (for example, as food) beyond just those exact features that the capacities were selected for. Developmental plasticity in turn may lead into direct adaptation to new environments (West-Eberhard 2003). Plasticity is especially important in humans. Yet another adaptive process to be taken into account is *niche construction* (Odling-Smee, Laland & Feldman 2003): the organism not only chooses the habitat but may have adapted to actively change its environment in ways that make the organism a better fit to it, and further evolves into those environmental products, making the organisms and their environments co-evolving interactive wholes.

All the above ways of becoming adaptive without an organism-changing adaption process (or Darwinian selection) are further

[63] There are, for example, species that do better in urban than natural environments, the most extreme example being the rock dove (*Columba livia*) – the feral pigeon. They did not evolve for urban environments, but they have found an ecological fit in these environments that is better than what they find in natural environments, after which certain features in their behaviour have further adapted to urban environments (Rose, Nagel & Haag-Wackernagel 2006). It is sensible to study the pigeon ecology as adaptively functional to cities even if only the finishing touches were produced by adaptation to this environment.

reasons to abandon empirical and historically explanatory adpaptationism – and they should be abandoned. At the same time, they address further needs, beyond the Darwinian adaptation process, for adaptive functionality analyses of traits. First, a description that includes a functional account of the trait's role in the overall form of life of the organism provides a deeper understanding of the organism, and the natural basis for such an account is from the point of view of what all organisms fundamentally do: stay alive, sustain themselves, and reproduce (see Wouters 1995; 2005; 2013). How an organism does this in its environment is simply a part of its biological description. Arno Wouters (1995) calls explanations referring to these features of the organism *viability explanations*: a trait is explained by how it satisfies an organism's needs. This is not a causal explanation for the trait's existence, but it is a functional explanation of how the trait contributes to the viability of the organism. Wouters even proposes that it is this idea of functionality that biologists use in their practices, never an etiological function, unless they are doing evolutionary biology specifically (Wouters 2013).

Note that "viability" only refers to the means to *survive* in the environment, and while the notion of "survival of the fittest"[64] has become the misleading slogan for natural selection, one does not have to be the fittest to survive, and natural selection goes beyond mere survival. Although *optimality* is only a theoretical endpoint of the adaptationist selection models and cannot always be expected to be found (for the reasons discussed above), natural selection can be expected to result in forms somewhere in between mere survival and optimality. And the other way around: natural selection is a partial explanation for viability and everything beyond. We do not need to restrict the considerations of adaptivity to optimality; adaptivity can be seen as a range of variation

[64] The phrase does not exist in the original *Origins* (Darwin 1859) – the origin of the term is Herbert Spencer (1864). Unfortunately, Darwin endorsed it and added it to later editions of the *Origins* – mostly because the expression "natural selection" has other misleading connotations, namely that of there being something that literally selects something. (See Gould 2002.)

from survival to optimality. Third, a functionalist description from an evolutionary perspective is still explanatory in a more minimal sense that I have called "ahistorical adaptation" above. Viability and adaptivity of a trait are more fundamental properties in the evolutionary explanations than the selection process: selection is based on the comparison between viability and adaptivity of individuals (or other units). I will expand on some of these points next, and I will discuss the third "basic concept" of function, the *adaptivity function*.

3.2.2. Adaptive Functionality and a Taxonomy of Functions

One of the main problems of adaptationism in all its forms is the presupposition of atomism (that is, each trait has a unique, isolated task it is selected for), when an organism is a holistic system (Gould & Lewontin 1979; Gould 2002). The organism's parts simply cannot be functionally independent, atomistic adaptations. At the same time, the whole must have some degree of adaptedness. We can switch from atomism to a holistic approach and take the overall adaptedness of the organism to be the starting point of a functional analysis. We can also approach the environment as holistic surroundings with multiple simultaneous adaptive challenges with some qualitative variation. If we analyse individual traits from this perspective, we end up with a vastly different idea of what counts as functional in the first place. This overall adaptedness, in turn, can only come from some sort of adaptation process (be it Darwinian selection, habitat selection, or something else) that is at least partly guided by the logic of natural selection. This is the case even if the parts of the adaptive whole have not evolved for the adaptive function they have now, and even if the adaptedness is not optimal. Moreover, even if a trait has evolved because of some other evolutionary factors, natural selection may sometimes be a part of an explanation of why it *remains*.

It seems therefore that ahistorical adaptation perspective has two explanatory functions. First, even if natural selection is *historically* only a partial explanation for the apparently purposeful design of the

whole organism, if one is interested in an organism's way of life in its natural environment, then the point of view of adaptedness (or functionality in environment) has an explanatory function. Second, ahistorical adaptation perspective has counterfactual power. If a change took place in the environment for which the organism is now fit, either there would be *some* change in the organism or its geographical location, or it would be doing worse than in the current environment, possibly facing the threat of extinction. There might be ecological consequences that loop back to the functionality of the organism in the changed environment. In this view, a particular trait needs not to have evolved for whatever its role is from the point of view of this adaptive whole; it is sufficient just to play a role.

This approach is not only a heuristic approach to the organism's design, but also **minimally explanatory***:* it identifies a selection pressure and a corresponding trait and has, therefore, explanatory power over a set of what-if-questions about possible alternative traits relative to which it *would be* selected for or against, as articulated above. The components of the natural selection mechanism with the trait's adaptive value included are in place, even if they were not included in the evolutionary historical explanation of the trait. Even if natural selection did not guide an evolutionary process in the past, it is an effective cause in the on-going process of evolution, and it is predictive of the future. To make an even stronger claim, this is the case even if developmental constraints prevent the relevant alternatives, and these constraints, instead of selection, are the cause of the persistence of the trait. This is because the selection pressures exist even then, too. This supports the counterfactual that if the developmental constraints broke, selection would still maintain the trait, selecting against the deviations. Of course, the explanatory *relevance* is even smaller in this second case than in the first, where exaptation is combined with natural selection that actively selects against alternatives. Nevertheless, they are both cases of over-determination for the conservation of the trait. We are not just referring to actual *historical* difference-makers here. Furthermore, the ahistorical adaptationist explanation is more

general than historical: it explains a type of evolutionary process, not just a single token (Potchnick 2007), and the knowledge of these generalities is important for evolutionary biology as an explanatory, not just a descriptive, historical discipline. Either way, simply pointing to the adaptive function does not explain the historical evolution of the particular trait.[65]

Philip Kitcher (1993b) has previously proposed an idea similar to what I have been sketching above. He has argued that causal role functions should be defined by their adaptive role from the perspective of the overall architecture or design. His aim is to unify the functional discourse of biology as follows. If an organism adapts to a new situation through a small change in its architecture, or even without any change, by acquiring a *new use* for some of its old parts, this new use should be the (overall) etiological function. The causal role functions of the parts that constitute this capacity are defined by their role in this (new) primary use (see also Buller 1998).[66] However, the Kitcher-style unification of function concepts fails as a way to understand all biological talk about functions (and therefore mechanisms). A biologist might have other legitimate questions that define functions (and mechanisms) (see Godfrey-Smith 1993). An obvious case would be understanding cancer (as discussed previously): we need to understand the parts of its development in a functionalist way without any reference to adaptiveness. Moreover, this approach has limitations even in functional biology, given that well-adaptedness is

[65] This issue is directly related to the issue of whether optimality models try to trace signs of adaptive histories and be "censored causal explanations" (Orzack and Sober 1994a & 1996; Potochnik 2007 & 2010) or if their use is about something completely different (Rice 2012). I will not go into this issue explicitly here, although I acknowledge that my subsequent discussion implicitly outlines an alternative account.

[66] For example, if long feathers in the front limbs of birds were an adaptation to capturing insects and a later exaptation for flying (Gould & Vrba 1982), they still have the causal role function of being a part of a device with an etiological function of flying. This is not a claim about their evolutionary history.

empirically false because of structural and developmental connections and constraints. Nevertheless, this perspective illustrates a way to understand the question about what a trait is for, and it outlines an adaptationist answer that is explanatory without being a historical claim.

There have also been attempts to define the function of a trait simply through its contribution to fitness without the trait needing to have evolved for this function, for similar reasons (for example, Ruse 1971; Bigelow & Pargetter 1987; Walsh 1996; Wouters 1995; 2005; 2013; Garson 2016). Dennis Walsh, for example, defines what he calls *relational function* as follows:

> "The/a function of a token of type *X* with respect to selective regime *R* is to *m* iff *X*'s doing *m* positively (and significantly) contributes to the average fitness of individual possessing *X* with respect to *R*." (Walsh 1996: 563.)

In other words, this is a function defined by what evolutionary significance the trait has for the organism (that increases its fitness in contrast to a range of alternative traits). I will call the family of function concepts similar to this the **adaptive function**.[67] This notion adds the

[67] There are different formulations. For example, Arno Wouters refers to viability instead of adaptivity, as mentioned above (Wouters 2005 & 2013). Justin Garson in turn defends a position he calls *generalized selected effect theory* (Garson 2013 & 2016). In this theory, the function of a trait is the activity that led to its differential persistence or reproduction in that population, as in the other selected effect function theories (such as Millikan 1984 and Neander 1991) but allows this cause for persistence to be something that the trait does over the lifetime of an individual. In other words, it does not need to have had a selection history; it is enough for the trait to have the effect now. Garson considers this a version of a selected effect theory that gets rid of the general problem of these theories – that many seemingly functional traits do not have a real function. His theory solves the problem, but since the "persistence of trait" over the lifetime of an individual can only mean that the trait makes a positive contribution to the fitness of an individual possessing it, this definition is a variant

idea of adaptivity to the concept of causal role function (or picks those causal roles that increase fitness) on the basis of current utility but does not require the trait's historical origin to be caused by this function (being an adaptation, having an etiological function). Unlike Kitcher's proposal, it is not meant to be a unificatory idea but an addition to plurality of function concepts. Its purpose is to explicate the concept of function such as the "function" in the practices of ethologists, which is understood as **current utility** (see Tinbergen 1963; Cuthill 2005; Sherry 2005).

Let us sum up the above discussion into a systematic view about the relation between adaptationism and functional explanations in biology. A charitable way to see the adaptationist tendencies in biological explanations would be this: it is a functionalist perspective on the overall design of the organism where the measure of functionality is the evolutionary adaptiveness. This functionality is descriptive but can also set the explanatory agenda. To clarify the different ways in which different adaptationist ideas are related to functionality, based on the above summaries of the discussions on adaptationism and the concept of function, we can build a hierarchical system of the concepts of function with a few additive dimensions.[68] The base concept (**F0**) is a *bare causal role* defined by consequences. This is not a function yet. There are four dimensions that can be added to make it a function:

F1 *the causal role in a biological system*, from the point of view of the purposive functioning of the whole organism, or maintaining the organism;

---

of adaptive function, not the selected effect function as an etiological function, in my categorization.

[68] Wouters 2003 provides a somewhat similar taxonomy of functions that starts with *mere activity* (my F0) and moves through *biological role* (my F1) and *biological advantage* (my F3) to *selected function* (which is a combination of my F3 and F4). Wouters's taxonomy is meant to be hierarchical, but these dimensions need not coincide.

F2 *the role in explanatory interest*, which may be, for example, a function in a mechanism that constitutes a "higher level" property (Craver 2013) and thus overlaps with the first dimension, but it may be abnormal functioning (in the sense of normality defined in the terms of the first dimension), for example a mechanism causing cancer, which means that it does not need to imply F1;

F3 *adaptive value*, which is not conceptually equivalent with F1, although, as I have been arguing for the past few pages, they tend to contribute to each other;

F4 *the function of the trait being (a part of) the explanation of the trait's existence*; this does not need to be an evolutionary loop, at least conceptually, but it is usually thought of as such.

F1 and F2 are two different types of causal role function, F3 includes adaptive functions, and F4 is the etiological function. The relations between 1, 3 and 4 are an empirical issue and the criticism of adaptationism is, basically, that they do not co-exist all the time. Furthermore, F3 (adaptivity) is not the only criterion for F2 (explanatory interest). In a historical adaptation explanation, F3 and F4 (origins) are combined. Thus, *atomistic historical adaptationist explanation* (the main target of the criticism of adaptationism) combines F2, F3 and F4 as a presupposition. A non-atomistic alternative, *holistic historical adaptationist explanation*, would likewise have all dimensions and would still need to presuppose a strong form of empirical adaptationism to be true. A non-evolutionary functional analysis combines F1 and F2. *Ahistorical adaptation explanation* combines F1, F2 and F3. Now, the next question is whether there is a need for an explanatory analysis that lies somewhere between historical adaptationism and non-evolutionary functional analysis.

3.2.3. A Case for Non-historical Explanatory Adaptationism

Evolutionary functionalist analysis (identifying what the given trait is *for*) is practiced in biology without any explicit reference to evolutionary histories. Although biologists probably often think there is a direct link to adaptation histories, which would make the practice a target of some of the criticism discussed above, this is not the only way to understand or justify the practice. What I have been calling "ahistorical adaptationist explanation" is a form of explanatory functional analysis in which the adaptive functionality (F3) is the chosen perspective (F2). This perspective is not a mere expression of personal or even disciplinary explanatory interest. As I have argued above, natural selection, and especially selection for being functional in the environment (either by Darwinian selection or by habitat selection) plays a special role in understanding the *overall architecture* and the *functioning of the organism in its environment* even if none of the specific traits can be presupposed to have any particular kind of evolutionary history – or even if the current habitat is not the one in which the traits evolved. Taking an ahistorical adaptationist position is to explain organism's parts as *being part of the design* that is fit to the environment and partly directed by natural selection (it being the only directional factor for fit). This involves identifying what the challenges posed by the environment are and how the organism copes with them, and this is a relevant point of view for understanding the organism and its behaviour in its environment even without any interest in its actual evolutionary history. Therefore, it is not contradictory for a trait to be a homology, for example, and the various species sharing the trait to possess different adaptive functions for it, as a part of somewhat different overall design.

Moreover, given that the natural selection is the only source of adaptively complex design, the overall relation between the organism and its environment needs to have some reference to adaptive functionality as an explanatory component. This is the case even if the overall design is sub-optimal: if there is any adaptive fit between the organism and its environment, this fit is a part of the organism's

overall functional description. But most importantly, if organisms really tend to be functional wholes with the capacity to interact with their environment in fitness-enhancing ways (and although this is an empirical hypothesis, I will boldly assume it to be true), even if empirical adaptationism is not true, a design stance with an adaptive perspective (Dennett 1995) will be a more productive source for testable and likely true hypotheses of the organism's anatomy, physiology, and behavioural tendencies than any alternative method. This entails the heuristic use of methodological adaptationism to be a plausible position.[69] This must be, however, a holistic approach that takes some of the suboptimal traits to be features rather than bugs, and information about the design should be drawn from phylogeny as well (see also Calcott 2009 & 2014). So, for example, whales having lungs instead of gills must be taken as a given design feature instead of a failure, after which the whale respiratory system can be reverse-engineered from the point of view of fitness maximization in its environment. Furthermore, the design cannot be simply assumed to be functional even in this narrower sense.

In short, a functionalist explanation answers the question about what the trait under explanation does that is functional in relation to the overall architecture of the organism, the environment, and the organism's form of life. It is a way to relate the organism to its environment. The explanation for *where this design came from* (or why there is sub-optimality) is a different, historical question. It is about evolutionary history. And this is where the explanations like developmental constraints and spandrels become alternative explanatory hypotheses for adaptationist hypotheses. The only presupposition that the ahistorical adaptationist functional analysis must make about historical

[69] I suggest this as a general principle, but like all such principles, there may be contexts in which it fails systematically. The criticism of adaptationism in the human evolutionary science controversy is not only about whether adaptation explanations are sensible, but also whether there is something in the human context that predicts systematic failures in prediction. I will return to this later.

adaptation is that it has played a significant role in one way or another for the overall architecture and its cohesion, and this is something that probably nobody would deny. But what I am proposing here, perhaps somewhat surprisingly, is that the importance of this perspective for understanding organisms is an argument for methodologically adaptationist research heuristics in non-evolutionary parts of biology, but not in evolutionary biology proper.

Consider an example. The European honey buzzard (*Pernis apivorous*) eats small mammals and birds, but it is also the only known predator of Asian giant hornets (*Vespa mandarinia*) (see Cocker & Mabey 2005). It hunts them in the following way. When it kills an animal, it leaves a small piece of meat on a branch where it can see the meat. When a hornet arrives and takes a piece of it, and subsequently returns to its nest, the buzzard follows it and eats the nest and the larvae. For some yet unknown reason, probably because of a chemical deterrent of some sort that the buzzard excretes, the hornets do not sting it. (The sting would be lethal for the buzzard.) This hunting behaviour is a complex pattern of behaviour that combines both general and specific characteristics of the buzzard's cognitive and behavioural capacities and physical characteristics. For example, its ability to follow the hornet, including perception and flying, are not specific to this behaviour, but the tendency to leave a piece of meat on a branch may be a specialized trait. This, too, might have had a function that was not specific to hornet hunting when it was evolving. Nevertheless, the hunting behaviour as a whole is an adaptation in the ahistorical, ecological and qualitative sense that it is something that makes if fit for its habitat. The parts of the orchestrated behaviour, as well as the physical and psychological traits involved in the behaviour, can be analysed as having an adaptive function in relation to the entire behaviour of hornet hunting, when this particular behavioural trait is taken as the adaptive complex under analysis. A part of our understanding of these various component traits and their role in the buzzard's form of life is in relation to this behaviour.

In other words, there is a form of behaviour that can be identified through it achieving something specific that has a positive fitness effect (eating the larvae) that binds the various participatory traits into one adaptive behavioural trait. The participatory traits constitute a mechanism that produces the overall behavioural trait, and their function, within this examination, is whatever causal role they play in the mechanism (that is, it is the Kitcher-style combined function). Some of the parts have other functions (most of them do), some do not (the hornet deterrent and setting the bait probably do not), but the task under scrutiny sets the analytical perspective for all the participatory traits and what their function is within this perspective.

This functional analysis does not say anything about the evolutionary origin of any of the specific component traits, or that this is the only or primary way to understand what any of the parts is for. They have many uses even from the perspective of adaptive functionality. Nevertheless, this is not just an analysis of causal roles in respect to the behaviour being analysed. Natural selection is likely to have played some role in building the hunting behaviour from whatever parts of it existed while this behaviour was evolving, tinkering with some parts, "exapting" others, and harmonizing the joint working of these parts in whichever ways possible. Complex behaviour like this could not be a pure accident, and probably nobody would deny natural selection to have played some role here. It is unlikely that complex multi-part traits would evolve otherwise, and if this is the case, the traits that take part in the overall behaviour do not only play a causal role, but their relation to the overall behaviour is such that they support whatever the trait does for fitness. If there is a complex behavioural trait, it is more likely that component parts, given the way the play together, are such that the overall behaviour is better for fitness of the organism than an alternative variant of the same behaviour. This is a functionalist perspective on the overall design and simply points out that when we slice the buzzard's doings in the forest into behavioural traits, it is reasonable to combine regular behaviour into larger wholes as adaptive tasks from the perspective of the fitness

effects of resulting behaviour. That is, it makes more sense to take "hornet hunting" as a functional design feature instead of merely categorizing the behaviour under structural descriptions (where external similar behaviour would the same behaviour regardless of the context).

We could perform the partitioning into traits in other ways, too: for example, we could be interested in a given cognitive capacity (such as observing hornets) and its overall function in all behaviour it is involved in, but this would be another explanatory question. Furthermore, we cannot understand the adaptive functions of cognitive traits without understanding what part they play in behaviour, since behaviour sets the fitness values for cognitive traits. Therefore, if we are interested in evolutionary explanations in particular, we must analyse behavioural traits as adaptive complexes and their constituent parts as having derivative adaptivity functions regardless of what they were evolved for. Most cognitive capacities have several functions (participating in several behavioural traits). It is likely, however, that these behavioural tendencies and the mechanisms underlying them co-evolve into coherently functioning wholes that support a multitude of complex behavioural traits like this, they have been preserved for their roles in these complexes, and the buzzard's behaviour and the ways in which it uses its environment (including the habitat of choice) have adapted to what it can do. However, this correlation is just the direction of causal influence from natural selection; it is not a guaranteed effect.

Furthermore, there is a selection pressure against changes towards dysfunctionality in all the traits involved in the hunting behaviour. When the behaviour has emerged, natural selection is a conservative as well as enhancing force.[70] The adaptive functionality is,

[70] A note on the term "force". Elliot Sober (1984) analyses evolutionary factors as forces analogical to Newtonian forces. This dynamical model has been challenged by the statisticalists that were discussed before (see especially Matthen and Ariew 2002), but also by some causalists, who consider it a misleading and potentially dangerous analogy (e.g. Lewens 2010). However, as Hitchcock and Velasco (2014) argue, much of the debate on this issue is based on an

therefore explanatory (in the minimal sense) for the existence of the hunting behaviour, even if the actual evolutionary history might mostly involve something else, which it undoubtedly does. Finally, the adaptationist perspective links various traits into an integrated whole, a complex trait. Explaining the complex behaviour involves identifying the component traits and their functions from the point of view of the complex trait and putting them into a mechanistic explanatory connection. This approach also predicts that many other, not yet studied traits (capacities, tendencies) to be such that they are more likely to advantage than to disadvantage the hunting behaviour.

For the methodological purposes of non-evolutionary biology, the usefulness of evolutionary functionalism such as this comes from the guidance the approach gives for "reverse engineering" the organisms. One can give a *structural* description of the organism's parts and capacities and what it does in its environment, but there is also a use for the partitioning and integrating that places traits in a *functional connection* with each other and with the environment. This leads back to the issue of whether we should only describe mechanistic structures and their operations or also their causal-functional composition. I have already argued for the latter, and in the case of behaviour, it is the only way we can make sense of the overall behaviour, given the parts that we are putting together (such as laying a piece of meat on a branch and eating the larvae) and the causal relations between them do not form Cartesian mechanisms of direct structural connectedness. Their connection in

---

inadequate understanding of Newtonian forces and the analogy as such is innocent. The terms "factor" and "force" are used almost interchangeably to refer to causal factors relevant for evolution, with the term "force" having a desirable connotation of continuous effectivity. This is probably partly because philosophers of biology do not share a common foundational theory of causation. Most of the work on these issues attempts to stay neutral regarding more foundational issues (although it is very difficult to do so, as I showed in the previous chapter). However, the terms such as "factor" and "force" are often used in a sense that attempts to be neutral regarding theories of causality.

meaningful ways can be a part of a discontinuous set of activities and separated by time.[71] Given a high degree of tinkering, organizing and guiding of evolution by natural selection, and given that the survival of the organism is tightly connected to the fit between its characteristics and the environment (be those characteristics evolved for this environment or not), a reasonable perspective for the functional description of an organism and its behaviour is what adaptive challenges it faces and how its characteristics meet those challenges. This is not to say that the traits have evolved for these particular purposes or that the architecture is perfect for those purposes that we are projecting onto them. But although this perspective is a projection, it is not mere projection. Hypotheses about yet undiscovered characteristics based on the projected purposiveness and already known architecture (including constraints) are more likely to be true than those that would go against purposiveness. The behaviour and design of an organism can be given various descriptions, but it is more likely that one guided by considerations of what the adaptive purpose of the functioning of the whole is, is a better heuristic for the discovery of new properties than any random description of what the organism does.

What cannot be presupposed, however, is that any *particular* part is functional let alone optimal, or if it is, that it has evolved for that purpose specifically. Furthermore, developmental constraints, historical contingencies, and the fact that each component part needs to function in relation to every adaptive task it participates in, should be taken into

[71] It would be tempting to think that the reasons why we undertake actions are the connectives for behavioural traits in the human context. The individual behaviours get their meaning from and are connected by reasons and intended goals for action. This is not available for all animals at least, and it is problematic in the human context, too, for explanatory purposes other than intentional action explanation. First, the identifications of reasons are not identifications of psychological mechanisms; second, the intentional action scheme applies to actions, not behavioural traits; and third, we may give evolutionary explanations for psychological mechanisms and behavioural traits, but not for reasons for action. I will discuss this topic in detail later.

account. The form of evolutionary functionalism I am proposing as the rationale for adaptationist research in biology is not a form of empirical adaptationism, but a tool for making hypotheses about the functional structure. Although it assumes that natural selection has played a central role in the evolution of the organism under study (for example, humans), it does not make strong assumptions about the strength of natural selection as a historical explanatory factor. It is analogical to making optimality models in evolutionary biology proper in order to discover not only selection but other guiding factors when the model fails in prediction. But what I am also claiming is that these two weakly adaptationist projects have different purposes.

As I said at the beginning of this section, ahistorical explanatory adaptationism like this could simply be called *evolutionary functionalism*. It is linked directly to both kinds of methodological adaptationism, evolutionary historical and ahistorical functionalism. Evolutionary functionalist analysis might be a way to discover the actual evolutionary history (along with the empirical evidence from, for example, comparative studies or fossil evidence). The use of ahistorical methodological adaptationism to discover the functional architecture of the organism, on the other hand, is an alternative to historical adaptationism to interpret the reverse engineering that Dobzhansky (1973) and Dennett (1995), for example, think is required to make sense of, and integrate, the different parts, functions and behaviours of biological organisms, as if the organisms were designed, as well as to integrate different subfields of biology – and, in Dennett's case, beyond. Both approaches use evolutionary functionality as a tool for discovering evolutionary histories or new traits in the architecture of, say, human cognition. But they are looking for answers to different explanatory questions and need to take different complications into account. Next, I will distinguish between three kinds of evolutionary functionalism as methodological tools.

3.2.4. The Kinds of Evolutionary Functionalism

To sum the discussion so far: there are three possible and meaningful but different goals to assign an adaptive function. The first is the **artefact model** of understanding biological systems (Lewens 2000 & 2004). In its strongest form it would slide into the form of empirical adapationism that Lewens (2009) calls "well-designism." But it is possible to have a weaker interpretation (call it *well-enough-designism*) and make only assumptions about the *overall* adaptedness to a relatively high degree. Some degree of sub-optimality can and must be assumed, and even a few instances of dysfunctionality every now and then, to make the ahistorical explanatory adaptationism a sensible perspective for a holistic causal-role functional analysis from the evolutionary functionality perspective (in the Kitcher style) – as long as there is enough overall adaptedness. As I have argued above, adaptive-functional analysis like this does not require an adaptationist evolutionary history. I propose this to be the charitable reading of the practice in biology to approach organisms from the evolutionary point of view, as if they were adapted, without committing the analysis to any of the problematic presuppositions discussed above regarding the criticism of adaptationism. This charity could probably be extended to at least some forms of evolutionary anthropology and psychology as well. The weaker form of artefact model is enough to make the practice of reverse engineering both a sensible stance for functional analysis and a source of hypotheses in non-evolutionary biology. But it *cannot* be used as *independent evidence* since anything may be non-adaptive.

It may be worth highlighting here that the artefact model, and the practice of studying biological entities from the design perspective (or from "design stance"; Dennett 1987 & 1995) in general, is always a *cognitive tool*, not a discovery of real design in the proper sense, *even when the apparent design really is caused by an adaptation process*. To start with a trivial point, there is no literal design in nature. The design perspective is an instrumentalist approach to make biological systems intelligible, although it is closely linked to the real patterns that

evolution seems to produce. There *seems* to be design in nature and we can analyse natural beings from this perspective. The existence of this design can be explained through real but non-purposive processes of evolution. Evolution's tendency to produce apparent design makes the practice of artefact model useful, but it is not a literally true description, just a cognitively salient way to understand the dependencies in nature.

But secondly, maybe less trivially, even true adaptation (as a historical process) is decoupled from the adaptationist design perspective as a guiding principle for describing organisms: the historical causes and the adaptive analysis in the current environment are two different things, *even if the environment is exactly the same now as it was then*. The logic of natural selection explanations goes the other way around. The adaptivity of the current use (in its environment) is explained by the comparative adaptive values and other factors in the past (in a similar environment). Otherwise, it would be a teleological explanation. Thus, the questions about adaptivity (or the functional analysis of the organism) and causal explanatory history are both conceptually and substantially separate even here. There is design that we *project* onto the organism, and we use its environment as a point of reference for what it was "designed for". And we acknowledge that this design is imperfect. The evolutionary history explains the structure of the organism that we perceive as functional. The history is partially an adaptation process to environments different but somewhat similar to current environments, even when the simplified adaptation story is true. What is needed for the artefact model to work is a reason to think that there is coherent complexity that could be functional in some environments, and a reason to think that the organism is functional in its current environment. The similarity between the environment of adaptive evolution and the current environment guarantees this, but it is not a necessary condition. Consequently, the adaptive-functionalist thinking in analysing an organism is not about evolutionary histories in any case. Natural selection having had a hand in the history is an explanation for why this thinking is somewhat

successful. The two (adaptive-functionalist thinking about a trait and giving the trait a causal-historical explanation) are conceptually distinct and they play different roles even in evolutionary historical explanations: observed adaptive functionality in the current environment may be used to formulate a hypothesis about similar environmental conditions having been causal factors in adaptive evolution in the past.

Therefore, it is sufficient empirical assumption for the artefact model to work that there has been some sort of selection-driven tinkering causing an apparent design in the architecture. This is enough to justify the practice. Every part of the design does not have to be evolved for precisely the purpose it is analysed to have. This happens at its fullest when the whole system under study is a true adaptation to a similar environment to the one in which it lives, but the causal explanation for the functionality of the design should not be confused with identifying design. This would be an *intentionalist fallacy*. It cannot be meaningfully considered to be a criterion for what is "really" an adaptation. The further question for historical explanations is whether we can use the *current use* and its environment as a clue to what the trait was selected for. And sometimes the adaptive design to be explained cannot be explained by processes that assume the similarity between environments past and present.

The second type of functionalist explanatory adaptationism is **functionalism in evolutionary history**, or in how the trait actually came to be the way it is. This is a more complicated issue. Even if natural selection were the most important evolutionary factor, for example in making sense of how apparent design is possible in the first place, it is still just one evolutionary factor among others that have a real significance. There seems to be a near consensus among philosophers of biology that all three dimensions of adaptationism (empirical, explanatory, and methodological) are fallacies in any strong sense when it comes to evolutionary historical explanations. Furthermore, the causal effects of natural selection change over time when both the environment and the population change: it is not a simple relationship between an environment and the selected trait. No biological

characteristic is entirely an adaptation to any single environment, strictly speaking. The causal contribution of any environment to any trait comes in degrees. The contribution of natural selection to evolution, as such, as central as it is, is an abstraction from a multifactor causal history. There is, however, a much more modest historical use for evolutionary functionalism in historical evolutionary biology. Adaptationist scenarios give *how-possible* explanations (Brandon 1990) that can be used to generate more specific follow-up questions.[72]

[72] The issue of how-possible explanations is quite complicated, however. It means different things in different contexts. William Dray introduced the idea of how-possible explanations as historical explanations that do not tell us everything that we need to know in order to understand how something happened, but are fragments of information combined with plausible speculation that *would* explain what happened. The goal is not, however, to generate a plausible hypothesis. The goal is to give an account of the events that shows that at least one plausible explanation exists that makes the event intelligible. It is not mysterious that the event took place. (Dray 1957 & 1963; see also Persson 2012; Reydon 2012.) Robert Brandon, however, defends adaptationist explanations as schematic explanations that have holes in them. They are *potential explanations*. (Brandon 1990; see also Resnik 1991.) These seem to be two different meanings of "how-possible explanation", and they have two different functions: Dray-type explanations are epistemic achievements in themselves that tell us why something was not unexpected, while formulating a potential explanation tells us what we still need to know in order to have an actual explanation (see Persson 2012; Kokkonen 2016). Furthermore, Peter Machamer, Carl Craver and Lindley Darden (2000) introduce the notion of *plausible explanation,* by which they mean an explanation in which the existence of a mechanism with certain causal powers is well established but the mechanism is a black box. This is a third form of how-possible explanation (Persson 2012; Kokkonen 2016). The confusion about different notions of how-possible explanation has been sorted out by distinguishing between *local* and *global* how-possible explanations (Forbes 2010) and between *possible causal scenarios* and *causal mechanism schemas* (Ylikoski & Aydinonat 2014). I have developed a hierarchical taxonomy of how-possible explanations and their different possible epistemic achievements elsewhere (Kokkonen 2016). Furthermore, I argue that a Dray-type how-possible explanation is an epistemic

Models based on them can also be used to discover the existence of other factors (see Orzack & Sober 2001b). This is only a tool among others and should be used jointly with hypotheses on other evolutionary factors (Lloyd 2015), but it is a central part of evolutionary biology. This also means that understanding the adaptive dynamics of, say, altruistic behaviour, is a large part of the evolutionary historical explanation of why altruistic behaviour exists wherever it does, even when other factors than those of the proposed dynamics (to be further discussed later) dominate the behavioural evolution.

The third type of evolutionary functionalism is simply **current utility**, or *adaptive function* without further inferences into proximate mechanisms or evolutionary histories. Describing this function has what I have been calling *minimal explanatory relevance*. It is *explanatory:* it specifies a part of the natural selection mechanism that is an active and present causal factor even if it was not a part of how the trait emerged historically, even if other factors are over-determining or frustrating its effects, or even if there has not been enough time for the mechanism to bring about the effect. It is also *relevant* as an explanatory factor. First, regardless of the actual evolutionary history of a trait (or our knowledge of it), its adaptive function tells us that this is the direction selection *would have* pointed to (and maybe it did) and it gives a direction for further evolution (given that there are no constraints). Second, it tells us that the trait would outcompete variants that are worse off for this particular use. Third, and perhaps most importantly, it also tells us about what kind of changes in the *environment*

---

achievement in the evolutionary context too. For example, we cannot possibly know the actual explanation for the origins of morality, but a plausible explanatory narrative makes morality understandable within the naturalistic framework. The further complications do not matter, however – the main point is that adaptationist scenarios have a heuristic function in evolutionary biology even without a strong commitment to adaptationism of any kind. The danger of sliding from innocent adapationist heuristics to full-blown empirical adaptationism is, however, a real threat (see Lloyd 2006 & 2015; Green 2014).

would create selection pressures for change and what kind of changes in the environment would be *detrimental* for the survival of the species (from the ecological perspective). It answers a set of what-if questions, and this increases our understanding of the organism in its environment. Nevertheless, the explanatory relevance is also *minimal*: it is neither necessary nor sufficient for explaining the trait's existence, nor, in the case of behavioural traits, for providing a correct functional analysis of the underlying proximate mechanisms. However, it can serve as a starting point in, for example, ethology for either speculation on evolutionary history (see Cuthill 2005) or for a functional analysis of the proximate design (see Bolhuis 2005). If so, we switch the topic. In the evolutionary context proper, we switch into the previous kind of evolutionary functionalism and should only consider the function as the first approximation, a *how possible* explanation that enables alternative hypotheses and the gathering of evidence, without discriminating between non-adaptive alternatives. If we use it as a starting point for a functional analysis of the proximate design, we switch into the artefact model, which, once again, should only serve as a heuristic tool for discovering the causal-role functional organization of the organism's physiology and behaviour, and how these interact with the environment.

Just as there are three ways to understand adaptive functions, there are three ways to understand evolutionary explanations as functionalist explanations. They are the following:

> **Current use functionalism**: the adaptive function is the context of analysis; minimally explanatory
>
> **Historical explanatory functionalism:** the adaptive function (current or past) also plays a role in the historical explanation of the trait's origin
>
> **Ahistorical explanatory functionalism:** the adaptive function is the guiding principle for the functional analysis of the organism and its behaviour

There is a close connection between these three forms of evolutionary functionalism but, as such, they address different explanatory questions. Adaptive functionality in the sense of current utility simply reflects what aspects of the trait are doing something positive for the organism, measured by fitness. Historical explanatory functionalism traces the contribution of adaptivity to evolutionary change. Ahistorical explanatory functionalism, or evolutionary functionalism in a non-evolutionary context, looks for functional explanations for what the organism is doing in its current environment. It is enough for the current utility function to refer only to the fitness effect, but both in the artefact model and in evolutionary histories adaptation is only one relevant perspective. There are other factors in the historical process of evolution and this matters for the artefact model as well.

I have been making a case for an instrumentalist approach to ahistorical evolutionary functionalism, but it is only justified by the fact that actual historical adaptive processes have a lot to do with the overall design of the organism. But this is not as controversial as true adaptationism, if it is controversial at all. The artefact model can accommodate adaptive features of the system that are not historical adaptations, but the process of adaptation as a whole is what makes the organism and its behaviour, in its totality, a functional whole to whatever degree it is functional. But this still does not mean that we should be happy with a functional analysis alone. Incorporating knowledge about constraining factors such as phylogenetic inertia, structural limitations, and path-dependencies makes even the artefact model more precise. For example, we should take whale's lungs as a central design feature of the whales even in the artefact model. Brett Calcott (2014), for example, has argued for such an approach, which might be thought of here as a *historically extended artefact perspective*.

Taking all these considerations into account, we may finally conclude that there are at least five different ways to interpret what an *evolutionary explanation* (not adaptation explanation only) of a given behavioural trait is:

1) a statement of the **current utility** (which is a minimally explanatory description),

2) a **causal-historical adaptation explanation** (with an etiological function),

3) an **ahistorical functional explanation** (with a causal role function, as a part of the artefact model description of an organism),

4) a **causal-historical non-adaptive explanation** (that is, a historical narrative with multiple types of causal factors, including but not limited to natural selection), or

5) an **extended artefact model** (which is not a historical statement but takes non-adaptive historical knowledge into account in its view of the design).

This distinction is important in interpreting the claims of "evolutionary social science" (the various schools of using evolutionary analysis tools on human social behaviour). Some of the evolutionary research programs seem to have different ideas about what an evolutionary explanation is in the first place – and by extension, implicit ideas about what is being explained. I will return to this in the next chapter. But first, a short note on problems that remain.

The three forms of evolutionary functionalism suffer from different adaptationist problems. The *current utility analysis* has no specific problems as such and is, in my estimation, a defensible adaptationist position, and ahistorically explanatory, but it is only minimally explanatory. It simply sets the trait in relation to environmental factors in a way that constitutes a part of the natural selection mechanism and tells us what the trait's use for the organism is, from the point of view of how it makes the organism fitter than (some of the) alternative forms of the trait would. This is not a fascinating achievement as such. The *artefact model* makes the additional assumption that the overall design of the organism is guided by this principle. In the weaker and more holistic formulation I have described here (in contrast to a Swiss Army knife style piecemeal adaptive

architecture), this form of methodological adaptationism does not require any of the parts under study to be true (historical) adaptations. This should be enough to use adaptive functionality as the guiding principle in reverse engineering the functional organization of the organism. This form of evolutionary functionalism cannot, however, be more than a tool for first approximation and generating hypotheses, given that dysfunctionalities are a real possibility, and the adaptationist ideas need to be amended. But this is also an instrumentalist approach to the design that gets judged by its practical usability. The use of *adaptationist heuristics in evolutionary biology proper* to outline the histories faces more serious problems, and I would conclude, as promised earlier, that adaptationism is not viable within evolutionary biology (for the reasons present in the literature and discussed above), even if it is a sensible ahistorical perspective on the design of organisms. Natural selection is only one and much constrained factor in the actual history, so even if it is the most important one, the use of evolutionary functionality as a guide to actual historical processes in evolutionary biology proper is misleading. Nevertheless, it may have a more limited use as a weaker *how possible* form of hypothesis formation, and adaptation explanations as such are central in evolutionary biology. It is important to understand *how* they work. This alone would make the clarification of adaptation explanation important.

The three forms of evolutionary functionalism have different epistemic functions and different limitations. The core of the analysis is the same, however, so I will discuss "evolutionary functional analysis" and "evolutionary explanation" without making further qualifications about which of these projects is involved. I will, however, assume that human sociality, at the level of generality I will be discussing, is a product of evolution, and natural selection has played a role in this evolution. The overall aim and motivation for the project at hand is so that we can give certain features of human sociality *evolutionary historical* explanations, and these explanations are part of what justifies the use of the *artefact model* in *understanding human sociality*. How far this methodology can lead us, in either project (the evolutionary origins of human sociality or understanding it in contemporary humans through the artefact model), is up to the

success (or lack thereof) of these research programmes, as discussed in the introduction. I will not return to the problems of adaptationism later, but this is not because I do not consider them significant. I will discuss some other complications, however: how assumptions about proximate mechanisms and developmental mechanisms affect how the adaptation explanation works.

I will move to the issue of how to relate the other explanatory dimensions to evolutionary dimension in the next sub-chapter. I will discuss adaptive function, evolutionary history, individual development, and proximate causes, using the famous four questions of Niko Tinbergen (1963). The main point will be this. Although there are different and, in principle, independent explanatory dimensions to any biological trait, how to answer them depends on what the answers to questions on some other dimensions are. One cannot apply evolutionary explanations or evolutionary functionalism directly to behaviour and leave the mechanistic basis for the behaviour and the developmental factors that shape the replication or reproduction of the trait as black boxes. These points are well established in the literature, usually in the context of arguing against adaptationism or making the case for the *Extended Synthesis* (Pigliucci & Müller 2010). I will argue later that the relation between individuals and the groups they belong to becomes more complicated in all dimensions (mechanistic basis, development, and evolution) in ways that might sometimes call for a holistic approach in any or all dimensions.

3.3. Tinbergen's Questions with Mechanistic Answers

Half a century or so ago, the ethologist Niko Tinbergen published an influential paper called "On aims and methods of Ethology" (1963), which, as the name suggests, discusses and distinguishes various aspects in the study of animal behaviour.[73] It distinguishes four areas of

[73] This distinction has since become the orthodox way to sort the different tasks of behavioural biology; see Manning 2005.

causal explanation in the study of behaviour. This distinction has become to be known as "Tinbergen's Four Questions" or "Tinbergen's Four Whys", because it distinguishes between four different senses in which a question like "why is the male sparrow singing?" can be understood. These senses are, according to Tinbergen:

1) **Causation**, which refers to the "behavioural machine" or mechanisms underlying the behaviour; for example: *What proximate causes make the sparrow sing?*

2) **Survival value** (almost invariably called "function" in more recent literature; see Hogan & Bolhuis 2005 & 2009), which refers to whatever it is that the behaviour does for the animal that makes it better off (directly or through indirect consequences) in its natural environment: *What is the singing for?*

3) **Ontogeny** (or *development*), which refers to how the behaviour and its underlying mechanisms appear in the course of the individual development of members of the species: *How did the singing behaviour come about and change during the sparrow's lifetime?*

4) **Evolution**, which refers to the actual history of the trait: *How did the singing behaviour come to exist on the population level in the course of the history of the species and its ancestors?*

As such, these questions are not precise enough to be four different *explanatory questions.* Rather, they are four different perspectives from which to frame explanatory questions about animal (and human) behaviour, mapping different *explanatory dimensions* (as I have been calling them), each of them collecting a set of explanatory questions about the same behavioural phenomena. Furthermore, each of them contains a set of questions that differ in more than just what kind of (complementary) explanatory information they seek. One could even go as far as to say that the different explanatory questions within a dimension

ask about different parts of the same causal network and the answers are either complementary or competing, whereas the different dimensions simply ask different questions (Sherman 1988). However, the questions in different dimensions are connected through the subject matter: the answers to explanatory questions in any given dimension may sometimes depend on facts about explanations given in some other dimension (Mitchell 1992 & 2002; Pigliucci & Müller 2010). This also means that the *formulation* of an explanatory question in one dimension may derive its motivation from needs in another dimension.[74] I will discuss the four explanatory dimensions (mostly) from the point of view of evolution of social behaviour now, taking the evolutionary functionalist stance explicated above to relate other dimensions to the evolutionary dimension. The evolutionary function is the perspective that sets the *relevant explanatory questions* in the other dimensions. The answers to these questions in turn affect the evolutionary explanation. This perspective is needed for evolutionary purposes, and it may be relevant for other purposes (for the reasons discussed previously), but this is not meant to be a perspective-free systematisation of how these questions are related.

3.1.1. Causation

Causation is a category that includes questions about both the external and internal immediate causal factors that trigger a certain behaviour, and about the mechanisms underlying these causal dispositions and participating in the causal processes that produce the external behaviour. Questions about external conditions only (*what proximate*

[74] In practice, questions that cross the boundaries between different research areas within biology are seldom in focus. When attention is paid to them, a whole new research program gets initiated sometimes. This has been the case, for example, in evolutionary developmental biology (Raff 1996; Hall 2003; Müller 2007). The boundaries within a discipline are not that different from ones between them sometimes.

environmental factors cause the behaviour?) leave the details of the "behavioural machine" black-boxed, concentrating on the interaction of the animal with its environment, to what stimuli it is responding in which ways. On the other hand, one could be interested in precisely this "machine". This interest includes both comparative-psychological (cognitive-ethological) level questions (animal cognition and motivational structures that are inferred from input-output relations in natural or artificial environments) and neurophysiological questions. These, again, differ in kind from each other, and both can be approached from both structural and functional perspectives. (See Hogan 2005; Shettleworth 2010.) A central question for the (proximate) causal explanation of behaviour is how to define what to explain: is the *explanandum* a singular behavioural occurrence (usually not) or a behavioural trait? If a trait, there are different ways to define it: all structurally similar behaviour (regardless of context), all behaviour in a given context, all structurally similar behaviour in a given context, or all behaviour that is produced by a given set of proximate causal factors. There is no unambiguous way to slice behaviour into traits in biology (see Lidicker & Freund 2009). What behaviour goes together as a trait depends on the explanatory interests. This is also the case with human behaviour (Longino 2013).

A further complication in explaining human behaviour is that we categorize both the behaviour and its causes as *action* guided by *reasons* and *intentions* in our folk-psychological practices. This practice individuates behaviour individualistically and (if given a causal interpretation) expresses its causes as well. However, if we approach behaviour from the evolutionary functionalist perspective (as we should for evolutionary purposes), we should instead approach behaviour along the lines that I discussed with the example of the honey buzzard. Various causes work in an orchestrated manner, forming a mechanism that produces a systematic behavioural pattern, and this pattern is identified through what adaptively significant effect it achieves. From this perspective, internal and external factors are connected: the evolution of internal dispositions (perceptual, cognitive, and motivational),

when guided by selection, is such that the combination of dispositions and external stimuli (in typical environments) jointly cause the behaviour that is functional in that environment. I will return to how to understand this in human behaviour and how this relates to our common way to understand human action in detail later.

A special case of environmental stimuli is those caused by other individuals. Social behavioural dispositions that result in (evolutionarily) beneficial outcomes *without* the individual processing the ends and means can be favoured over dispositions that involve individuals actively processing the ends and means. In human context, these dispositions would include things like emotional reactions to other individuals' behaviour. They do not need to determine the behaviour to be selected dispositions; learning and reflection can shape both the perception triggering the emotion and the range of behavioural responses within the emotional state and the social context while the emotional response is still an evolved biasing factor for behaviour (see Mallon & Stich 2000). Non-reflective dispositions are likely to evolve if the social settings are robust enough to couple social triggers and a functional reaction reliably, since individual processing (contra selected reactive dispositions) needs to go through proxies (that is, individuals need to want things that lead into fitness-increasing behaviour, and they need to recognize these things in a situation; see Sterelny 2003), processing consequences takes time and effort, and it is prone to errors that may be unnecessary. If there are persistent or frequent features in the social environment, selection favours reactive dispositions over individual cognition – although the two may be mixed. The selection for social environment does not need to be different from evolution guided by other features of the environment: behaviour is a selected response to the environment as it is, guided by the cognitive and motivational mechanisms that are selected to bring about this behaviour under the (social) environmental conditions, based on the clues that individuals are able to perceive within the environment and to process.

An individual does not need to *aim at* those consequences that are beneficial, but simply achieve those consequences because of what she or he is aiming to do: people may want to achieve something, but given the constraints, they achieve something else, which is what they actually "need" from the evolutionary point of view (see Sterelny 2003). For example, people do not have an evolved urge for beneficial amount of energy intake, but for things like sugar, which lead into a healthy energy intake with availability constraints of the environment of adaptation, while urges without these constraints may lead into too much energy intake. This holds in social environments too. We may want to help others, for example, because unbeknownst to us this benefits us (in an evolutionary sense) in the long run. An unavoidable difference from other conditions is that the selection of social behavioural traits is dynamic: the fitness of a trait depends on other traits in the population, making the evolution sensitive to the repertoire and frequency of other traits in the population. But this is still a competition between traits.

However, the fitness-relevant effects of the trait may depend on the response from others in a way that the behaviour makes sense only as a *relational trait* between two or more individuals: the consequences of the behaviour depend on the form of interaction, not just the individual disposition to participate in the interaction in a particular way. This means that the *explanandum* for an evolutionary explanation of such behaviour is non-individualistic. This, of course, depends on both the facts about the proximate mechanisms guiding the behaviour and the adaptive functionality. I will make a case for this later. Nevertheless, the selection process itself may still be individualistic selection between individuals, with mutually beneficial consequences emerging. Game-theoretical models of behavioural strategies are usually interpreted in this way. I will argue that it may be inadequate to consider the *object* of evolutionary explanation (even if game theory is used in modelling the behaviour) to be only *individual* behaviour or psychology, but sometimes the structures of social interaction may need to be taken as the object of explanation. The idea is roughly the following.

The behaviour has positive fitness effects for the individual because of indirect consequences that depend on the behaviour of other individuals. If you are comparing the fitness of individuals in the population, it is not enough to know their individual behavioural dispositions, but also what kinds of social interactions they are involved in. The object of the evolutionary explanation should be the collection of causal factors responsible for the link between the behaviour and its fitness consequences. If this collection is robust enough to evolve, and if it is a system of several individuals, the evolutionary explanation will be holistic in the proximate dimension. It is, however, another question whether the *evolutionary* explanation itself needs to be holistic (that is, the selection takes place between groups of individuals) or if it can still be individualistic (that is, the selection takes place between individuals only). I will discuss this issue in the last chapter.

3.1.2. Ontogeny

Ontogeny, or *development*, also includes both internal and external factors. In this case, however, what counts as internal is not so clear. Sometimes, if the gene's point of view is adopted, the distinction is made between genes and what counts as environment for their functioning – including the hormonal environment within the organism. There are two possible justifications for this. First, the idea that genes are the source of intrinsic information that, in a linear causal sequence in the development, produces definite effects on phenotype, in interaction with the external factors. This image, however, has proven to be wrong: development involves, even on the molecular level, not only coding sequences of DNA (the molecular genes), but also regulatory sequences, the RNA and protein products of the DNA's functioning, and the environmental signals that influence the regulatory machinery in the cell, which in turn is a functioning living cell from the very beginning (see Wolpert *et al* 2010; Griffiths & Stotz 2013). An alternative way to divide between internal and external is to start with the fact that the development of an organism is a process, in which

some causal factors are external to the *process*, and some are internal to it. In this, internal factors include both the genome and the non-genetic organismic properties in the developmental processes that build on the previous steps of the process. Even then, some of the external factors may be causally essential to development: some of the internal factors of development depend on some external factors such that they are not "fundamentally" internal. This, however, is a more realistic image of the process. (Oyama, Griffiths & Gray 2001; Griffiths & Stotz 2013.)

The second argument for the first way to make the distinction would be an evolutionary-theoretical argument from the gene's point of view *a la* Dawkins (1976): the genes are the developmental resources that reproduce and are there necessarily, in contrast to what depends ultimately on environment. This is an *ad hoc* distinction from a purely developmental point of view, but if the question comes from the evolutionary point of view instead, the very interest to make distinctions is founded differently. The perspective could entail different criteria for what is explanatorily adequate. This solution, however, does not work since it makes unwarranted assumptions about reproduction even from the evolutionary point of view, including modularizing the evolutionary units of reproduction and assuming only one route for the reproduction of traits (see Oyama, Griffiths & Gray 2001; Jablonka & Lamb 2005; Godfrey-Smith 2009; Pigliucci & Müller 2010). Furthermore, the concept of gene used here is an abstraction of inheritance patterns from the actual developmental processes, not something that could be identified with the molecular genes that are being copied and are the ones that matter for the developmental process (see Moss 2004; Griffiths & Stotz 2013). I will return to this in more detail later.

The external factors include the environmental factors external to the organism that participate in the developmental processes, both during embryonic development (the mother being the most relevant part of the environment) and after that. They may include specific properties of the environment that the development is sensitive to during the special periods of sensitivity for those particular factors, as

well as more generally affecting behaviour-changing properties. The factors of the developmental process can also be divided, for example, into more and less fundamental factors (for example, through the depth of *generative entrenchment*; Wimsatt 1986 & 1999), and into *permissive* and *instructive* factors (Woodward 2001; Griffiths & Stotz 2013; Calcott 2017). Both may include internal and external factors. Molecular genetics is the study of one and only one component of this. I will return to the discussion on whether it is illuminating or not, or even seriously misleading, to differentiate between internal and external in development in the first place, especially in the development of behaviour. But even though molecular genetics can be said to study a component of individual development of behaviour, much of behavioural genetics does not: it studies the association between genes and traits on the population level and should be classified as a different, although not independent, category. Evolutionary approaches work on the population level, and development used to be a black-boxed process between population genetics and population-level phenotypic variation, but this is an oversimplification. (See Wolpert *et al* 2010; Oyama, Griffiths & Gray 2001; Carroll 2005; Griffiths & Stotz 2013; Dediu 2015.)

The role of the social (including cultural) environment that affects development is an important special case. As with proximate mechanisms, the social environment can be just a part of the (developmental) environment from the evolutionary point of view, or it could play a more specific role. Individual development always takes place in the interaction between internal and external factors, but what are important in this from the evolutionary point of view are the individual differences that are transmitted to the next generation: the offspring resemble the parents in the ways that their parents are dissimilar from other individuals (Godfrey-Smith 2009). The features of the environment (and of the genome as well) that are not difference-makers for this, but are instead shared, indifferent regarding the outcome (even if they contribute to development), or random, do not matter. The factors that have the capacity to transmit traits, and differences in them,

to the next generation (or that participate in their reconstruction in the next generation; Griffiths & Gray 1994; Oyama, Griffiths & Gray 2001) can be targets of selection. This transmission may, at least in principle, take four different routes: genetic, epigenetic (that is, biological factors transmitted from mother to the offspring other than DNA that affect the developmental process), behaviour copying, and symbolic learning (Jablonka & Lamb 2006).

From the evolutionary point of view, individualism and holism in the developmental dimension can be defined as follows. If the reproduction or reconstruction of the traits under selection relies only on internal factors (genetic or otherwise) regarding the systematic similarities between parents and offspring and the differences between individuals, these differences being preserved between family lines (that is, the transmission is only *vertical*), the evolutionary process is individualistic in the developmental dimension. This leaves several options open for environmental factors in development: they may participate in the causal processes but the variation in these factors does not affect the outcome; there may be a range of variation that depends partly on the environmental factors in a non-discrete way; or there may be a limited number of discrete variants where specific environmental factors decide which of the possible variants gets developed. The object of evolutionary explanation may be a variant of a trait, but it may be the existence and frequency of the alternatives on the population level as well.

A charitable reading of the so-called *nativist evolutionary psychology* (more about this soon) is that it takes the object of the evolutionary psychological explanations to be those aspects of psychology that may have a range of variation but for which the range of variation is explainable as an adaptation. The alternatives may be explained by factors in individual development but the range of developmental dispositions by evolution alone. A charitable reading of what a trait being "innate" means would likewise work along these lines. I will go deeper into this issue later. I will defend a limited usability of the concept of innateness, and I re-interpret nativism as a form of

methodological individualism. Nativist evolutionary psychology is narrowing its explanations to those aspect of psychology that develop because of internal factors only. I would, however, consider linear behaviour copying from parents to offspring and even difference-making linear cultural transmission as forms of individualistic transfer of traits. The route of transmission does not make difference from the evolutionary point of view.

The holistic approach is needed when the reproduction processes are not only external but take place in larger settings. Human developmental environment consists of cultural products and meanings, and social practices and roles that require multiple individuals, and these elements are shared by groups of people. If the transmission of behavioural traits relies on multiple individuals contributing to their development, the group stabilizes the forms of behaviour collectively, and this is partly *horizontal* (not within family lineages alone), the transmission is holistic, group-level process. The aspects of behaviour that are transmitted in this way bind the adaptedness of the individuals together, as well as enables group-level adaptation. The key here is that the difference makers are shared – the same logic applies to "shared genes" in the inclusive fitness and to culturally- or behaviourally-transmitted shared similarities.

3.1.3. Evolution and Survival Value

Evolution concerns the change in the trait in the population over evolutionary time. It includes two components: building a phylogenetic tree (which was Tinbergen's main question about evolution) and telling the history of how the trait gradually changes in the course of evolution. The phylogenetic tree can be built through comparative studies (especially using genetics) and a speculative history of the trait's evolution can be sketched using the phylogenetic tree and (partially) adaptationist scenarios based on our knowledge of the evolutionary environments. In addition to this, *population genetics* is a perspective on evolution of its own that studies the process of evolution from the

genetic perspective. Furthermore, explanation using population dynamics differs in kind from a casual-historical explanation of how a trait has changed in the course of evolution. Brett Calcott (2009) has made this point and labelled the latter as *lineage explanation*.

The category of *survival value*, or *function*, is inherently ambiguous. It is meant to capture the ecological aspect of the behaviour: what is the role of that particular behaviour in the environment of the animal from the perspective of what positive feedback it has for the animal in its environment (see Krebs & Davies 1997; Cuthill 2005; Shettleworth 2010). Tinbergen's formulation equates function with survival value, making it what has been discussed as "adaptive function" above. As such, an ecological description of behaviour like this does not require it to be optimal in the environment or to be historically adapted for the environment, just to have a *use*. But there is no need for the trait to be adaptive at all: if the aim of behavioural ecology is to describe the animal's behaviour in relation to its environment in general, there is no reason to exclude those behavioural traits that are even disadvantageous in relation to the environment, any more than there is a reason for a physiologist to exclude those physical traits that are disadvantageous from the description of the animal's physiology. These traits cannot be said to have function, but there is no reason why ecology should be constrained to functional aspects alone. On the other hand, some ethologists seem to think that for it to be sensible to talk about function in the first place, it has to be a historical adaptation, while current use is only a guide to this (Cuthill 2005, for example, is explicit on this point). The function in behavioural biology can be understood, then, in all the senses discussed in the previous section: adaptive function, causal role function, or etiological function (that is, adaptation).

What I would suggest, however, is that the best way to understand the ecological analysis of behaviour through its survival value is the artefact model in the moderate sense I discussed in the previous section: an instrumentalist perspective on the overall design of an organism and its behaviour. If it is the animal's ability to exist in the

given environment we are interested in, which is what ethology is mostly interested in, then constraining the primary research questions to these aspects makes sense. The adaptation perspective is needed to make sense of the overall life form of the organism in its environment. Its recurrent behavioural patterns are bundled and sliced into traits from this perspective, and they are given a functional description from the point of view of their adaptive function in this overall image. The evolutionary source of the organism's behaviour is not an issue: it may be a combination of adaptation, exaptation, habitat selection, and plasticity of behaviour and its development (including individual and social learning, even culture in some species). These may all produce enough adaptive functionality to make adaptation to environment a sensible starting point for ecological analysis even without stronger (historical) adaptationist commitments.

3.1.4. Interdimensional Connections

The explanatory questions of different dimensions seek explanations for qualitatively different things. The answers to them cannot be combined into a unified grand story: even if they all refer to the same overall network of biological processes, they partition this complex net in different ways, being interested in different types of causal questions. However, as Sandra Mitchell (1992 & 2002) has argued, the explanations are still not, strictly speaking, separate: they may be relevant to each other by constraining each other's presuppositions and they may therefore clash indirectly (see also Pigliucci & Müller 2010). The lack of unification does not entail independence.

I have characterized the proximate and developmental dimensions from the evolutionary perspective above, but only to articulate which explanatory questions are relevant in each dimension from the evolutionary perspective. What I will do in the reminder of this subchapter is give an outline of how they are related. These connections are a reason for why individualism and holism (in the sense defined here) in the proximate and developmental dimension are relevant for

individualism and holism in the evolutionary dimension, and how they are relevant. The nature of the connections is mostly an empirical question but finding out the answer to this involves philosophical issues, some of which I will discuss in the rest of this dissertation. What follows now, as a prelude, is a discussion on the basic logic of how these dimensions relate. This logic may be characterized as a *mechanistic model of evolutionary explanatory hierarchy*. This hierarchy is not global: it is not about theoretical importance, causal primacy, fundamentality, or any such thing, nor a universal taxonomy of biological explanations. This is not an attempt at theoretical unification either, but a framework for how different explanatory dimensions are related, in accordance to the mechanistic analysis presented earlier, when evolution and evolutionary explanation is what we are interested in.[75]

Even if the evolutionary functional analysis of behaviour were understood as an instrumentalist artefact model without a need to presuppose *historical adaptation* of the particular trait, there would still be a presupposition of an *evolutionary process* producing the fit between the behaviour and the environment, guided by natural selection to a considerable degree, as discussed above. Furthermore, the connection between the behaviour and its adaptive function presupposes two other kinds of mechanistic connections: psychological (or neurological) mechanisms connecting the stimuli of the environment to the behaviour, and the developmental mechanisms guiding the ontogeny to produce the selected mechanisms for the selected behaviour in each generation. In this way, the evolutionary study of behaviour is a study of internally linked mechanisms (in the liberal sense of mechanism argued for in the previous chapter) in which mechanisms are components for other mechanisms – in other words, different *mechanistic levels* (in the sense of Craver 2007). In this case, the

[75] An account based on evolutionary functionality might be developed to understand how the different fields of biology are related to each other, as discussed in the introduction. The model I am about to give could be a first approximation for this. The aim here is much more modest, however.

mechanistic levels are not ontological levels or levels of biological organization, and the explanatory hierarchy here is not an ontological hierarchy in the part–whole sense or indeed any other ontological sense, but only in the explanatory sense.

In short, the adaptivity of behaviour requires an evolutionary process (guided by the mechanism of natural selection) as a mechanism to bring about the developmental mechanisms that produce the proximate mechanisms that in turn produce the behaviour. There are no "different kinds of causation" as such in this picture (such as "ultimate" and "proximate"[76]). There is a historical process with a causal

[76] The well-known distinction between ultimate and proximate causes, put forward by Ernst Mayr (1961; see also Ariew 2003), in which "ultimate" refers to evolutionary causes and "proximate" to everything else, is often mapped onto Tinbergen's questions scheme simply as "ultimate" equalling "evolution" + "function" and "proximate" equalling "causation" + "ontogeny". This equation is incorrect. First, "ultimate" cause is meant be an answer to the *why*-question, referring to evolutionary origins, and "proximate" cause is meant to answer the *how*-question through an answer describing the biological mechanisms on various levels, but development may be both proximate and ultimate in this sense (see West-Eberhard 2003; Laland *et al* 2011 & 2013; Calcott 2013). Secondly, if ultimate explanations are meant to give explanations through survival value from the point of view of population dynamics (i.e. "the trait *had to* evolve to that direction, and here is why"), there are further problems. If the functionalist analysis of behaviour is understood as an artefact model, it seems to fall into this category (for example, Gardner 2013 identifies "ultimate" with "adaptive rationale"), but it is not evolution. On the other hand, evolutionary histories without population dynamics would probably not qualify as ultimate explanations (Calcott 2009 & 2013). Furthermore, if proximate causes are meant to be individual-level processes, where do behavioural genetics belong to? It also seems that the ultimate–proximate distinction is not precise and one-dimensional in the first place: it seeks to capture the differences between evolutionary and non-evolutionary, population-level and individual-level, as well as adaptivity as a quasi-teleological equilibrium and directly causal processes. These do not bound together theoretically or empirically. Continuing to use this distinction hinders progress in understanding evolution rather than promoting it (see also Ylikoski & Kokkonen 2009 and Laland *et al* 2013). The main point of the distinction is that there are two kinds of causal processes in biology, but clearly, there are many, and they are

mechanistic structure that can be broken down in different ways, such as the Tinbergen's Questions scheme, resulting in qualitatively different causal explanations or narratives. One narrative is the description of the evolutionary history of a given behavioural tendency as a population-level process with natural selection and other factors. The environmental challenges and the behavioural solutions are abstractions on the population level, and the totality of the evolutionary factors can be conceptualized as a mechanism that produces the behaviour (in the evolutionary sense). However, this process presupposes an actual evolutionary history, which is a process of interactions between individuals of the population and the environment, and reproduction. This process can be partitioned causally in different ways. Discovering the details of the proximate mechanisms and their development can be approached as opening important black boxes in the mechanistic explanation of the evolution of a particular behavioural trait, and the information about them may be important for the population-level evolutionary explanations.

If things were simple (which they are not), the four explanatory dimensions would form a neat hierarchy of explanatory levels. I will now present this simplification to articulate the basic logic, which is the basis for adding complications.

---

related in a way that is made more sense of by a richer taxonomy and a mechanistic model of their relation that is being presented here now. At the same time, the very terminology of "ultimate" and "proximate" lures us to think that causes might be "more or less" ultimate or proximate and to conflate ultimate with distal, which almost contingently is one of the differences between evolutionary- and individual-level explanations. This might be congenial with the statistical view of evolution, in which all real causal processes are on the same level, but not with other views. I will be using the term "proximate" strictly in the sense of "proximate mechanism" in the category of causation in Tinbergen's scheme, except in chapter 5 and 6, where I discuss these mechanisms specifically. There I will use the terms "proximate" and "ultimate" to refer to different dispositional properties within the motivational architecture. The difference between evolutionary "purpose" and causal-mechanistic cause is, however, an important distinction as such.

1) *Adaptive function* determines the course of *evolution* (that is, understanding how the pieces that form the mechanism of natural selection fall together in the given context is all the mechanistic understanding we need of the evolutionary process), which
2) determines the *developmental pathways* (what gene-environment interactions there are), which
3) determines what kind of *proximate mechanisms* there are, which
4) determines the organism's *behavioural response to the environment*.

In this picture, the evolution of behaviour could be thought of as natural selection fine-tuning the intermediate mechanisms in a way that produces functional behaviour, and all transitional mechanisms between adaptive function and behaviour could be black-boxed. This would lead into an image of the study of social behaviour that is associated with Wilsonian sociobiology (Wilson 1975), where behaviour is modelled as directly adapted to the environment and the adaptive function should be the primary focus of behavioural biology, everything else being details.

This image is a simplification, but not entirely wrong. It articulates the rough lines of how the dimensions are related and why the idea of evolutionary functionality could be a viable way to integrate biological phenomena. However, every step in the scheme is more complicated. First, as already discussed, there are other factors in evolution: historical adaptationism is a fallacy. Even if the artefact model is useful in the ecological analysis of behaviour, and even if it were a useful heuristics to make inferences between adaptive function and proximate mechanisms (see Bolhuis & Verhulst 2009 and Shettleworth 2010), the evolutionary history including non-adaptationist factors is a more efficient guide for a heuristics in discovering the architecture of the proximate mechanisms. Even this has shortcomings and needs to be done in interaction with all directions: the integration of the

different dimension needs to be integrative, not reductive (see Bolhuis 2005; Cuthill 2005; Ryan 2005). This does not make behavioural ecology based on artefact model or adaptive functionality useless in analysing ecological relations, but it leaves the explanatory aspect of the analysis minimal (in the sense discussed in the previous section) and diminishes the heuristic value of adaptationist assumptions.

Secondly, there is the issue of *behavioural plasticity*. The proximate (neural or psychological) mechanisms of an adult do not produce the same behaviour to the same environmental triggers in all cases. (That would be a reflex.) There can be much greater sensitivity to specific overall contexts, creativity, and learning from experience. The biggest issue, however, is the role of the development of these proximate mechanisms in the first place.[77] Environmental factors contribute to it, which means that changing environments may affect the outcome (in many ways, including learning in a much deeper way than merely reacting to previous experiences). Sometimes this *phenotypic plasticity*, the capacity of a genotype to produce several different phenotypes, is a selected property of development.[78] Furthermore, the development constraints the variation and possible evolutionary tracks for the trait, binds together parts and functions of an organism that would seem like separate traits from a structural or functional perspective, and so on. The theoretical integration of evolution and development that is under way right now may still change our understanding of evolution

[77] As a historical note, the other three of Tinbergen's questions had previously been explicated by Julian Huxley, and the ontogeny dimension was added by Tinbergen partly because of the criticism of oversimplifying the developmental process. (Bolhuis & Giraldeau 2009.)

[78] By this I mean the phenotype on the level of proximate mechanisms. These mechanisms themselves can be sensitive to the environment in the sense that the same phenotype under proximate mechanisms description can produce different behaviour depending on the information available. Behaviour being a part of phenotype, this is also phenotypic plasticity, strictly speaking. But it may be useful to make the distinction between the plasticity of the development of mental faculties and the plasticity of behaviour with these faculties.

substantially (see Oyama et al 2001; West-Eberhard 2003; Piggliucci et al 2006; Sansom & Brandon 2007; Laland et al 2014). Evolution requires the transfer of traits to the next generation and the nature of this mechanism matters. *Evolution is not the evolution of traits, but of the developmental processes that produce traits*. The nature of these processes is relevant to the relation between evolution and the traits, in numerous ways – but it is not a reason to dismiss the general logic of the relation between the explanatory dimensions implied by Tinbergen's questions as such.

It is also important to know how development is related to proximate mechanisms and what is the role played by the developmental environment. The same environment can play different causal roles (Brandon 1990): it may be a part of the selective environment in the evolution, a set of causal factors in the development (in both ancestral and current population[79]), and a set of proximate triggers for behaviour. In all cases, it may participate in the mechanisms (proximate, developmental, or evolutionary) that produce the behaviour (in the proximate, developmental or evolutionary sense of "production"). These environmental factors may include the social behaviour of the other individuals and culture. The issue regarding individualism and holism arises in all dimensions, and holistic elements in one dimension can affect how the explanations work in another dimension. This depends on the adequate description of a relevant mechanism (in whichever dimension).

---

[79] Even if the ancient developmental environment (which is also the selective environment) is similar to the current developmental environment (in which we study the behaviour), they need not play the same role in the development. The role of an environmental factor may change in the development in the course of the evolution of the developmental mechanism of which it is a part. An extreme example of this would be the Baldwin effect, where the development of the trait gradually becomes insensitive to the environmental factors that were originally essential to the development, even a source for individual *learning* processes (Baldwin 1896; Futuyma 1998).

This leads us back to the issue of what the trait under evolutionary explanation is in the first place. It is not certain that the same thing can be considered as a trait in all the dimensions. In the evolutionary functionalist framework, behavioural occurrences that serve the same purpose (or are related to the same ecological challenges) are bundled together as a behavioural trait, as discussed before with the example of the honey buzzard. We may also be interested in the cognitive capacities and motivational mechanisms as such. They may participate in the production of several such behavioural "bundles" or behavioural traits, but from this perspective, these mechanisms are the traits, connecting all the resulting behavioural dispositions. This means that the evolutionary explanations of behaviour and of psychology may be explanations for traits that do not structurally map onto each other. Yet they are intimately connected. A behavioural disposition is selected for its role in the overall behaviour. If the behaviour is selected, the underlying proximate mechanisms are selected, which in turn are selected for all their behavioural consequences. And conversely: behavioural traits are bound together by the underlying proximate mechanisms that are, in turn, bound together by the behavioural traits that they participate in. I will return to this later.

Development adds a further layer. Shared developmental resources (be they genes or environmental factors) and processes may connect structurally and functionally unconnected proximate mechanisms together. Since the development of traits is what is being selected instead of proximate mechanical structures directly, this more holistic bundle of proximate- and behavioural-level traits is the unit of evolution and should perhaps considered to be the trait from an evolutionary point of view (see Gould & Lewontin 1979; Lewens 2009). This is so especially if the perspective is evolutionary history and phylogeny, not adaptation. Common ancestry leads into similar developmental processes that lead into similar traits and these could also be units to individuate traits for some evolutionary purposes (see Griffiths 2006). However, even if developmentally defined traits, rather than phenotypically defined traits, are the objects of selection, the

adaptive value that determines the selection pressures is still due to the interaction between specific phenotypic characteristics and the environment (although the developmental aspect is also relevant to this interaction), and this is the focus here. The developmental issue of consequence for the main topic is the role that the environment plays in development and the possibility that those developmental factors are an alternative route for the replication of traits.

To sum up, there are three important issues related to the explanatory hierarchy presented above. First, what is the logic of the relationship between the different kinds of biological explanation? I have sketched an outline for a general mechanistic framework, but I will expand the discussion on the specific points in the forthcoming chapters. Second, within the context of the explanatory hierarchy, is there a one-to-one mapping between what counts as a trait on different levels? I have raised the possibility of this not being so in the above discussion and I will return to this issue later. Third, are there explanatorily non-reductive, non-individualist elements on any of these levels in the case of human social behaviour? If so, what are the consequences of this for its evolutionary explanation, taking the first two issues into account? I will discuss the various explanatory dimensions in turn. Before this, I will give a brief overview of the human evolutionary sciences and show how frame these questions appear in the context of actual evolutionary work done on these issues.

4. Evolutionary Human Social Sciences

The human species is a product of evolution and has been studied as such in different ways. An obvious object of study is human evolutionary history. Another way is to use evolutionary ideas in the study of contemporary humans. Humans have been studied as biological beings from both perspectives, but sometimes evolutionary models and ideas have been used as analogies, metaphors or expansions from the properly biological sphere to somewhere else (such as economics or cultural phenomena). The interest here lies in the nature of evolutionary principles used in explaining sociality (for example, cooperation) and culture. I will concentrate in the "natural" phenomena; I will not discuss evolution within created systems (such as economics or science). Sometimes the evolutionary approach is central to the very identity of a discipline, sometimes evolutionary considerations are just an additional source of hypotheses, but this distinction is irrelevant as such for the current discussion. I will use the labels of "sociobiology", "evolutionary anthropology", and "evolutionary psychology" in my discussion, but these titles are meant to be inclusive – that is, although some of the discussion in the section of "evolutionary psychology" will be on the research program that is self-titled in this way, I include evolutionary approaches in "traditional" fields of psychology in this category.

These labels are not meant to capture different "schools" of evolutionary human sciences either – to capture the whole variety of approaches, more fine-grained categorization would be needed. The various approaches are not necessarily in direct competition, either, but rather ask different questions with explanations that may sometimes be compatible and sometimes not (see Smith 2000: Ylikoski & Kokkonen 2009; Brown *et al* 2011; Laland & Brown 2011; Brown & Richerson 2013). Furthermore, I will not evaluate any evolutionary approach as such here. The purpose of this chapter is to build the contextual background for the subsequent chapters: it matters, for the evolutionary social sciences, whether the explanations and the presuppositions

they make are individualistic or holistic, on all dimensions, and a failure to pay attention to holistic aspects of social behaviour when needed is a methodological failure. I will also introduce some key ideas and explanatory principles that will be discussed in more detail in the subsequent chapters.

4.1. Sociobiology, Broad and Narrow

There are two different senses in which the term "sociobiology" is understood. In the broad sense, it simply means all study of animal sociality and social behaviour from an evolutionary functionalist perspective (see for example Alcock 2003). In this sense, the object of this dissertation is sociobiology. In the narrow sense, it refers to a particular school of doing so (call this *classical sociobiology* if you will), which happened to be the school that started this line of research. Edward O. Wilson (1975) was behind the name "sociobiology"[80] and was responsible for the first systematic collection of its methodological tools and for building a unified theory, but the tools as such came mostly from theoretical biologists such as Robert Trivers and William Hamilton. The two central ideas of sociobiology were the *gene's eye perspective* on adaptivity of behaviour and the use of *game theory* to analyse social situations, both of which have become fixed features of evolutionary approaches to social behaviour. Both approaches are usually considered to be individualistic approaches, although they can be argued to be neutral in this regard – Elliot Sober and David Sloan Wilson, for example, have argued that some cases in which these tools are applied are cases of group selection (Wilson & Sober 1994; Sober & Wilson 1998; see also Okasha 2006 and Birch 2017). Furthermore, the sociobiological tools make no assumptions about the proximate or developmental mechanisms at all. I will now review some key aspects of

[80] The term itself was introduced as early as the 1940s, but it did not gain a fixed reference and did not see any particular use before Wilson (Plotkin 2004: 105).

classical sociobiology, and also some other tools of sociobiology in the broad sense.

4.1.1. The New Synthesis

The study of behaviour has always been a part of biology, but the pioneers of modern animal behavioural biology, *ethology*, with evolutionary component as one of its central parts, were Karl von Frisch (1886–1982), Konrad Lorenz (1903–1989), and Nikolaas Tinbergen (1907–1988), who shared the 1973 Nobel Prize in Physiology or Medicine for this.[81] The approach became an established research area within zoology by the early 1950s. As we have seen, the ethologists thought from the very beginning that there are several biological ways to approach behaviour, including evolutionary approaches. The biological approach, generally speaking, was expanded to comprehend human behaviour by ethologists in the 1960s and 1970s (for example, by Konrad Lorenz (1966 [1963]) and his student Iranäus Eibl-Eibefeldt in a more academic context, and by Desmond Morris in a series of popular books; see Laland & Brown 2002; Plotkin 2004). This *human ethology* challenged the culture-centred view of many anthropologists of the time, replacing it with a zoologized view of humans. At the same time, they acknowledged the many peculiarities of human behaviour (such as culture) and their own methodological limitations in studying humans properly (Laland & Brown 2002, 59–64)). However, early ethological work influenced some anthropologists, including Lionel Tiger, Robin Fox (for example, Tiger & Fox 1971), and Donald Symons (1979).

Human ethology was arguably based on evolutionary functionalism, guided by the idea that animal behaviour can only be

[81] Their studies, of course, built on existing tradition, starting from Charles Darwin himself (Darwin 1872), with most important prior advances arguably made by Oskar Heinroth, Charles Otis Whitman, and Julian Huxley (Burkhardt 1981; 2005).

understood in natural environments – laboratory experiments are inadequate or at least insufficient – but it was not restricted to study of function. Ethologists were also interested in proximate causal mechanisms and developmental processes without an evolutionary perspective, and much of their evolutionary attention was in building phylogenetic trees, not in discovering adaptive functions. (See Laland & Brown 2002; Plotkin 2004.) The tools of sociobiological analysis started to appear during the 1960s and 1970s (for example, Hamilton 1964a, 1964b & 1970; Maynard-Smith 1964; Trivers 1971 & 1973; Maynard-Smith & Price 1973). Edward O. Wilson collected the methods in a systematic theory in his field-coining book *Sociobiology*, subtitled *The New Synthesis*, referring to the expansion of the "old synthesis" of Darwinian evolutionary theory and Mendelian genetics into the modern evolutionary biology (Wilson 1975). Wilson's *Sociobiology* and Richard Dawkins's *The Selfish Gene* (1976), which is a more philosophical take on the guiding principles of the new approach and its gene-centred ontology, popularized the approach inside and outside academia. The approach pushed proximate and developmental questions and phylogenetic considerations into the background, concentrating on the function of behaviour. The first human applications of the approach appeared by the late 1970s, amid fierce controversy (see Segerstråle 2000). In retrospect, leaving the proximate and developmental questions aside was an obvious step for the worse from the more traditional approach of ethology, and this was partly guided by mistaken ideas such as very strong adaptationism (Gould & Lewontin 1979; Kitcher 1985) and unwarranted behaviourism (which also reflected the differences in approaches between European and American psychological as well as zoological traditions; see Segertråle 2000; Laland & Brown 2002), but an important motivation behind sociobiology was the emergence of new theoretical tools, such as kin selection and game-theoretical models, to understand social behaviour from the evolutionary point of view. A charitable reading of this new approach

is that its proponents got a bit overexcited about the new methods and approaches.[82]

The most central theoretical idea of the approach was the gene's eye view, or *gene-selectionism*, which switches attention from the fitness of individuals to the fitness of genes. Although this shift of attention is *away* from individuals, it was built on direct criticism of group selection models. The early proponents of group selection (for example, Wynne-Edwards 1962) thought that the adaptation of behaviour is sometimes *for the good of the group*. The individualists (for example, Maynard-Smith 1964 and Williams 1966) pointed out that even if individuals behaving for the good of the group make the whole group fitter, behaviour that makes the individual fitter than other individuals in the group (where all individuals get the benefits of being in the good group) will still be selected. This means that individual benefits trump group benefits every time they are in conflict. The logic in the gene-selectionism, however, is that it is the genes of the individuals that matter. The genes that are associated with behaviour that makes those very genes better, *no matter in which individual*, is selected. If a gene is associated with behaviour that promotes the fitness of the copies of the same gene in other individuals, it can be selected.[83] One crucial aspect of this move is to distinguish the different functions of the *copying entity* and the *phenotypic entity* in the logic of the selection process, as explicated by Dawkins (1976; see also 1982) and elaborated by David Hull (1980, 1981 & 1988a). The copying entity (*replicator*) is the proper carrier of fitness, while the phenotypic entity (Dawkins's *vehicle*, Hull's *interactor*) interacts causally with the environment in ways that determine the fitness of the replicators. I have already discussed some complications to this idea, raised by the complexity of developmental processes that are left black boxes in this image. The whole

[82] Furthermore, as I explained in the previous chapter, black-boxing proximate, developmental and evolutionary-historical mechanisms makes some sense, although this was a mistake nevertheless.

[83] What exactly "gene" is referring to here is a more complicated question that will be returned to later.

approach has been since questioned (for example, Griffiths & Gray 1994; Oyama, Griffiths & Gray 2000; Godfrey-Smith 2009). I will return to the problems and the remaining insights of this distinction later.

4.1.2. Kin Selection

The key notion of gene-selectionism is the concept of *inclusive fitness*. It was introduced by William Hamilton (1964a & 1964b), based on previous work by Ronald Fisher (1930) and John Haldane (1932 & 1955). Inclusive fitness consists of two parts: *direct fitness*, which refers to the positive effects the trait has on the organism's own reproductive capacity, and *indirect fitness*, which refers to the positive effects the trait has on the fitness of other organisms who share the same genetic basis for the trait. Both promote the selection of the underlying genetic basis that the trait is connected to. This is the basis for the effect that John Maynard Smith (1964) coined *kin selection* and Robert Trivers (1985) proclaimed to be the most important idea in theoretical evolutionary biology since natural selection itself: helping your kin to increase the fitness of your own genes. Formally put, when

> a_i is the direct positive fitness effect the trait has on the individual i,
> b_{ij} is the positive fitness effect i has on the individual j,
> c_{ij} is the negative fitness effect on i for having the effect on j,[84]
> r_{ij} is the multiplier from the degree of relatedness between i and j, and
> w_i is the inclusive fitness of i,

the inclusive fitness can be calculated from the equation

$$w_i = a_i - c_{ij} + \Sigma\, r_{ij}\, b_{ij}$$

[84] The behaviour may be costly, in which case there is a loss in *absolute fitness*. But even if there is no such loss, merely helping someone has a negative fitness effect on *relative fitness*. I will return to this (and to this distinction) later.

where sigma refers to the sum of the fitness effects from all relevant individuals. The value of the relatedness multiplier is the same as the degree of the shared genes (when only those genes in relation to which there is variation in the population in the first place are considered) on average – in diploidic organisms, the multiplier between the parents and offspring, as well as between siblings, is 0.5, between first cousins 0.25, and so on. These multipliers, however, presuppose that the parents are not related at all, and the population is infinite. Relaxing these unrealistic idealizations decreases the multiplier. In a population consisting of only one family of two generations, the multiplier would always be zero. According to *Hamilton's rule*, which is the central equation of kin selection, altruistic behaviour of i towards j can get selected, if $rb - c > 0$, that is, if $rb > c$. (Futuyma 1998, 595–596.)

I will return to kin selection later, but two observations should already be made here. Relatedness as such is not a causal factor. What is important is that the gene to be selected is associated with behaviour that benefits individuals who have that same gene, which includes the individual themselves and some others. The explanatory power of Hamilton's rule does not come from relatedness, and even less from the *overall* shared genome, but from the *likelihood* of the gene in question existing in the other individuals, which is higher the more closely related they are. There is a *correlation* between the *amount of shared inheritance due to relatedness* and the *probability of sharing the gene*. The first is not explanatory, which makes "kin selection" an unfortunate phrase.[85] Another important point is that the two causal mechanisms in kin selection (the direct fitness benefits and the indirect fitness) are very dissimilar factors. Direct fitness has to do with the individual's reproductive capacity compared to other individuals, whereas indirect fitness makes sense only through the structure of the population and therefore requires a different mechanistic explanation even if the two factors could be modelled in the same equation. This has one

[85] This is not exactly so straightforward, but it is sufficient for now. I will discuss kin selection in greater detail in the final chapter.

particularly important consequence. Although sociobiology was interpreted as an individualist account of social evolution by its proponents and critics alike, this is not the only possible interpretation of these tools. Even if genes were the only replicators that matter (Dawkins 1989 [1976]), both individuals and groups might still be the realizers of the interactions that the genes are selected for – in other words, the vehicles or interactors (Wilson & Sober 1994). Hamilton himself, for one, thought kin selection was a form of group selection (Hamilton 1970) after being convinced by George Price (1970) on this point. Elliot Sober and David Sloan Wilson (for example, Wilson & Sober 1994; Sober & Wilson 1998) have made a strong argument for this, too. I will return to this issue in more detail in the final chapter of the dissertation.

4.1.3. The Evolutionary Game Theory

The other central theoretical element in sociobiology is the use of *evolutionary game theory*, developed especially by John Maynard Smith (Maynard Smith & Price 1973; Maynard Smith 1982). Social interaction can be modelled as a game where different options for behaviour have different fitness consequences depending on how the other(s) in the situation behave. The games are *repeated* and the number of rounds is unknown, since the focus is on the behavioural traits (or patterns of behaviour), not individual instances of social behaviour. The overall fitness of any behavioural disposition depends on the frequency of encounters with the various types of behaviour, and in an evolving population (with heritable behavioural dispositions) the dynamical selection process (that is, the very selective environment that evolves in the process), too, can be modelled as a series of games – this addition of the dynamic aspect is what differentiates evolutionary game theory from classical game theory. It is used specifically to understand conflicts and cooperation, but not only that. The central idea is that if the formal system has an *evolutionarily stable strategy* (ESS; Maynard Smith 1983), that is, no other alternative strategy can replace it, the system will ultimately reach it. The strategies are obviously idealizations,

not descriptions of actual behavioural traits, and they concentrate solely on the emerging fitness effects. Furthermore, the games are about consequences, not what the "players" are "aiming for" in the situations. This means that a social setting may have a certain game-theoretical structure measured in fitness without the concrete social interaction (defined with the individuals' preferences, for example) having an isomorphic game-theoretical structure. The illustrations usually simplify this by concentrating on context where something concrete with a fitness consequence (for example, resources or other direct consequences of the behaviour) is "in play". The structure of the evolutionary game may correlate with the structure of behavioural-level aims, but not always, and these structures cannot be collated. Still, they are useful tools for understanding the dynamics in interactions.

There are different games to capture different social situations, but the classic of this approach (and the one that will be examined later) is the formalization of Robert Trivers's (1971) theory of *reciprocal altruism*. The starting point for this is the intuitive idea that mutual help makes the individuals participating in the interaction fitter than individuals who do not participate. The possibility of *free riding* (taking help and not reciprocating) should, however, make such tendencies unlikely to evolve. This situation is a classic case of *prisoner's dilemma*. If we have only *defecting* and *cooperative* strategies, it is always better to defect, no matter what the other player chooses to play.[86] However, if reciprocity is an option, it may be the winning strategy. Reciprocation can be modelled as a *Tit-for-Tat* strategy (TFT), where the player cooperates first, and then reacts to however the partner played the previous round – keep cooperating if they did, otherwise

[86] For example, if cooperation costs one unit (of whatever has robust effects for fitness) but gives three units of benefit for the partner, defecting against another defector leaves you with zero units, whereas cooperating would only cost you a unit, and defecting against a co-operator pays three units, whereas cooperating would only grant two units. The only ESS in this setting is defecting with zero gain, even though everyone cooperating would give two units for everyone.

defect. This strategy does worse than defecting on the first round, but if there are enough reciprocal players and/or pure co-operators, it may do better overall. If it invades the population, it becomes ESS.[87] (Futuyma 1998, 584–586; Hargreaves Heap & Varoufakis 1995, 197–198; Sober & Wilson 1998, 79–80.)

Once again, the game-theoretical models have usually been interpreted as individualist models, and understandably, so – they model individuals and their behavioural strategies in social situations. But, once again, there are alternative interpretations in which game-theoretical models are sometimes interpreted as descriptions of group-level selection processes (see Sober & Wilson 1998). Furthermore, it is not so clear what the abstract strategies and games are modelling in the real world. I will get back to these issues, too.

4.1.4. Group Selection

Not all early models of social evolution were individualistic. David Sloan Wilson had already presented his *trait group model* during the emergence of sociobiology (Wilson 1975), although much of the foundational theoretical work and its popularization was done much later in the 1990s, notably as a collaboration between Wilson and Elliot Sober (Wilson & Sober 1994; Sober & Wilson 1994 & 1998; see also Sober 1980a & 1984 and Wilson 1989). Other pioneers of both theoretical and empirical work include, for example, Charles Goodnight and Lori

[87] TFT does not automatically do better than defecting – it does not necessarily invade the population. But if it does, it is able to become fixed. There are also a lot of ways to make reciprocal strategy better in realistic ways, given some cognitive capacities. For example, in *observer-TFT* the player chooses the strategy of the first round based on previous observations about the partner's strategies (Pollock & Dugatkin 1992), and in a variation of this the player uses others' attitudes as a clue for this (Castro *et al* 1998). In *strong reciprocity* (Gintis 2000a) the defectors are punished – this is a problematic case, since even if this can make defecting unfavourable, punishing may be costly. (This may be a reason why, for example, lions tolerate free-rides.)

Stevens (Goodnight, Schwartz & Stevens 1992; Stevens, Goodnight & Kalisz 1995; Goodnight & Stevens 1997). Much of the later discussion on individualism and holism has been about what *counts* as group selection – whether kin selection and the evolution of reciprocal altruism are really forms of group selection, for example, as mentioned above (see Sober & Wilson 1998; Okasha 2006; Goodnight 2012; Birch 2016). Wilson's trait group model, however, is an unambiguous case of group selection in which the group level differences direct selection and override the invasion of selfish phenotypes from within. Even if a selfish (individual-benefitting) type is always fitter than the altruistic (group-benefitting) type within every group, all individuals in a group with more altruists are fitter, which may make altruists fitter in the population overall. If there is a mechanism that keeps this structure constant, group selection takes place. However, the issue of individualism and holism (as defined in the introduction) is partly about how we should interpret the sociobiological (in the broad sense) models as *models of causal mechanisms in evolution*. The fitness consequences that follow certain patterns (specified by the models) are a necessary condition for selection but modelling only the consequences does not say anything about the mechanisms that cause these patterns. This is a matter of relevant causal processes on all three levels discussed: proximate, developmental, and evolutionary. Group selection will be the topic of the final chapter.

4.1.5. Biological Markets

The *Biological Markets Theory* put forward by Ronald Noë and Peter Hammerstein (1994, 1995 & 2016) is a major new development in the evolutionary modelling of social behaviour. This theory approaches some forms of interaction within the species (such as collaboration, trading of goods, and mate selection) but also between species (such as cleaning mutualism or pollination) as "markets". The idea is that there are goods (such as food or gametes) and services (such as cleaning or warning calls) that can be treated as commodities and the

individuals involved can be treated as traders. Traders choose their trading partners based on supply and demand, and on "bartering value". Markets are formed only in situations where commodities are traded voluntarily, multiple potential partners are available, and there are differential bartering values. The theory also includes the idea of "advertising" services by signalling the services, which might include false information.

Biological markets seem to work along two dimensions. The "currency" of trading in abstract is fitness, and the evolution of behaviour is explained as an evolutionary scale "bargaining process". This does not need to involve any bargaining or partner choice in the proximate dimension. For example, the evolution of symbiosis between flowers and pollinators can be formulated as a biological market, but the behaviour of pollinators is simply a result of this process. On the other hand, some animal behaviour looks a lot like actual bargaining; for example, primates exchanging grooming for other services seems to be a case of using grooming as a currency in actual trading. I suggest that we distinguish these two aspects of biological markets: evolutionary biological markets and proximate biological markets. At the same time, evolutionary biological markets explain the proximate biological markets and why the value of the currency of the latter can be measured in fitness. The idea of biological *markets* sounds like a metaphorical use of the term "market", but if the market model applies to the evolutionary context and identifies a similar mechanism, the situation is analogical to the use of natural selection models outside biology: regardless of the disciplinary context of the discovery of a mechanistic structure, the structure itself may be more common. The ontological status of such repetitive structures is not important here. However, it is possible that real economic markets are an outgrowth of an activity that itself is a biological market in the proximate biological sense, and that this is only one case of biological markets in humans. I will not go deeper into this here, but the approach has been applied to humans as well, outside economic activities (see Barclay 2013 & 2016).

The Biological Markets Theory is about reciprocity, but it is not a variant or extension of the reciprocal altruism model – their explanatory mechanisms are different. They are not necessarily in competition either. However, they may be competing explanations for any particular case (see Carter & Wilkinson 2013, for example). Both can be used to explain reciprocity in behaviour (that is, altruism in turns over time), but whereas reciprocal altruism is about interaction between individuals, biological markets are about choosing the partner for interaction. The two mechanisms may contribute to the evolution of interaction at the same time. The distinction will be important later.

4.1.6. The Shortcomings of Sociobiology

There are different ways to interpret the evolutionary functionalist analysis of behaviour, as I have been arguing above. Classic sociobiology (or sociobiology in the narrow sense) concentrated only on the function of behaviour, but its aim was not only an ecological analysis of current utility. It might be possible to reinterpret classic sociobiology as a form of ahistorical explanatory functionalism, but it was presented as historical explanatory functionalism that was ahistorically explanatory as well, because of strong adaptationist assumptions that, as we have seen, provoked the classic criticism of strong adaptationism. To be fair, however, the sociobiologists did not claim that the other dimensions (proximate and developmental) or other evolutionary factors and actual histories are *causally* irrelevant, only *explanatorily* irrelevant. The other factors could be black-boxed, since natural selection is the only directional factor in evolution (even if it is not the only one) and since only the genetic material that builds the behavioural capacity is inherited and therefore persistent (Wilson 1975: 551; & 1978: 56 & 172; see also Laland & Brown 2002: 95–101; & Alcock 2003). Furthermore, Edward O. Wilson (1975: 551) believed that species-level traits evolve rapidly under new conditions, while only traits general to higher taxonomical levels are robust enough to resist rapid change. All this became subject to criticism of *genetic determinism* (for

example, Rose, Lewontin & Kamin 1984; Kitcher 1985) and adaptationist story-telling (e.g. Gould & Lewontin 1979; Kitcher 1984),[88] and these assumptions became especially problematic when sociobiological methods were applied to human behaviour.

One way to express the shortcut that classic sociobiology makes between behaviour and its evolutionary function is that it presumes the hierarchical mechanistic scheme that I presented in the previous chapter. The logic is as follows. The actual evolutionary processes have a causal function in the mechanistic process of natural selection, but their details do not matter. Other factors in evolution are not important in the long run. Evolutionary processes, in turn, involve the transfer of traits from one generation to the next. Developmental processes are the mechanisms that produce the behavioural capacities and tendencies in a complex interaction between various factors, but only genetic components, being the only inherited parts of this, are relevant to the evolutionary functionality, and they, too, can be simply assumed. Behaviour, of course, involves capacities and tendencies, that is, proximate mechanisms, that cause the organism to behave in the selected ways, but the details can be left black boxes as well. Much of the justification for all these black boxes, in a charitable reading, comes from the *population-level perspective*: only certain aspects of the processes are robust enough to have a population-level effect.

[88] To be fair, the sociobiologists were not genetic determinists in any direct sense (see Wilson 1978; Segerstråle 2000), and even Kitcher (1985) only criticizes them for being indirect genetic determinists, since the sociobiological thesis was never that genes determine behaviour, but that other developmental factors do not matter to the evolution of behaviour, since they are largely irrelevant to heritable differences between individuals. For this to work, patterns of heredity should be seen in the diversity of human behaviour. The roots of behavioural genetics lie in eugenics, and it was Jerry Hirsch who brought the population genetics to ethology decades earlier (Greenspan 2008), but the need to prove heritability of human behaviour made a link between classic sociobiology and *human behavioural genetics*. But population-level genetics leaves the development inside the black box.

Therefore, the adaptive function determines (to an explanatorily relevant extent) the evolution, which determines which evolutionarily relevant developmental factors (genes) exist, which then determine the proximate dispositions, which determine the behavioural outcome – and so the adaptive function is all we need to know to understand the behaviour.

None of the black-boxing is warranted, and all the mechanistic details matter. Contemporary evolutionary human scientists agree with this. They incorporate the sociobiological *methods* with more substantial ideas about the mechanisms that connect the behaviour to its evolutionary function. Some of them concentrate on mind (*evolutionary psychology*), whereas some are still attempting to analyse the behaviour and its context, including culture (*evolutionary anthropology*).

4.2. Evolutionary Psychology

Like "sociobiology", the expression "evolutionary psychology" can be understood in several ways. Broadly speaking, it can refer to any approach to psychology that is informed by evolutionary considerations, be they historical or functionalist. There are, broadly speaking, three types of approaches (that are not mutually exclusive): evolutionary histories of mind, evolutionary functionalist methodology within psychological research, and evolutionary psychology proper. Evolutionary histories of mind try to trace the natural historical development of human mind and its capacities, although usually with the aim of informing non-historical psychological research as well. Examples include relatively broad phenomena like the theory of mind and other social and communicative capacities (Byrne & Whiten 1988; Corballis & Lea 1999; Brüne & Brüne-Cohrs 2005; Tomasello 2009 & 2014), morality (Bekoff & Pierce 2009; Boehm 2012; de Waal *et al* 2014), religion (Boyer 2001; Atran 2002; Bering 2006; Schloss & Murray 2009), or language (Carruthers 2002; Mithen 2005), and these approaches usually combine evolutionary functional considerations with comparative

empirical data from humans and other primates (or even larger clades), and are usually more interested in the temporal sequence of the emergence of new capacities (and the related phylogenetic issues) than functionality (see also Gangestad & Simpson 2007). Both this kind of evolutionary speculation and evolutionary functional analysis of psychological capacities and tendencies are sometimes used in theorizing or as heuristics in psychological research that does not attempt to reveal evolution as such, but to learn about how the human mind works (for example, Baron-Cohen 1995; Narvaez *et al* 2012). These two types of evolutionary inquiries into mind (evolution of mind and evolutionary methods in psychology) are separated by their aims, but they are connected in much of the substance. The third type I mentioned above, "evolutionary psychology proper", includes approaches that take the evolutionary perspective (historical or functional, or both) as the basis for understanding the workings of mind (and sometimes human phenomena beyond just mind), and they practically combine the aims of the two other types directly (for example, Barkow *at al* 1992; Buss 2005 & 2014), usually in an (historically) adaptationist manner.

The classification of these approaches does not matter as such, and I do not evaluate any specific approach in this dissertation.[89] But

[89] For the criticism of the boldest evolutionary approaches, see Buller 2005; Richardson 2007; Ylikoski & Kokkonen 2009; and Smith 2020. Two aspects have to be distinguished in the problems commonly associated with the field of evolutionary psychology: the deeper problems having to do with theoretical and methodological assumptions and more superficial problems having to do with quality of research. Much of the epistemically bad reputation of the field may be related to the latter, and even in this case, the issue is not so much the quality of the research performed as it is the structural constraints under which it is done. First, evolutionary psychology shares the general problems of psychological research. The notorious *replication crisis* is probably symptomatic of several methodological problems (such as lack of actual replication of the experiments and a high level of theory-ladenness), but probably includes the overestimation of the uniformity of human psychology (see Henrich, Heine & Norenzayan

it is worthwhile to note that "evolutionary psychology" is not one thing. Furthermore, the main idea for the study at hand is what evolutionary psychological approaches say about human sociality. Even here the focus is only on what difference the assumptions about individualism and holism make – and that these assumptions exist. To articulate this, I will briefly describe the methodological starting points of the evolutionary psychology movement that claimed the name of "evolutionary psychology" first, and is also known as the Santa Barbara School, nativist evolutionary psychology, or "Evolutionary Psychology" (with capital letters) (see Laland & Brown 2002; Buller 2005; Sterelny 2007), to highlight some clearly individualist tendencies. After this I will review some criticism and alternative takes that may provide reasons for the holistic perspective.

2010). Evolutionary biology, on the other hand, always suffers from a high degree of speculation and uncertainty. Evolutionary psychology accumulates the problems of these fields. Furthermore, whereas evolutionary approaches to animals can assume that the current habitat of the animal is approximately the same as the environment of evolutionary adaptedness, in the case of humans, this environment (including its social aspects) is speculative. To make things worse, the idea that the mind that evolved in this environment and that the minds of contemporary societies are the same, or highly similar, is an assumption that falls outside empirical evidence. All this makes evolutionary psychology highly speculative and vulnerable to assumptions that originate in preconceived ideas about "human nature" and the social interactions experienced by the researchers, as well as their value basis. This is all familiar criticism that has already been directed at sociobiology (Kitcher 1984; Segerstråle 2000). My point is, however, that even if the epistemically bad reputation of evolutionary psychology is deserved, this does not mean that evolutionary approaches to mind suffer from these problems *necessarily*, or that the deeper problems are actual problems. Furthermore, these theoretical and methodological assumptions may be correctable.

4.2.1. Nativist Evolutionary Psychology

Nativist evolutionary psychology (or *nativism*) is a branch of evolutionary human science that was born partly out of the criticism of classical sociobiology and was critical towards it too. Its main founders were Donald Symons, Leda Cosmides and John Tooby, in the late 1980s (Cosmides & Tooby 1987; Tooby and DeVore 1987; Symons 1989; Tooby & Cosmides 1989), and other prominent pioneers include Jerome Barkow, David Buss, Martin Daly, and Margo Wilson (for example, Barkow *et al* 1992; Buss 1995; Daly & Wilson 1999; see also Laland & Brown 2002, 153–157). According to Cosmides and Tooby (1987, 278–279), the failure of (classical) sociobiology was to try and explain behaviour directly (as genetic adaptation to present conditions), when it should be obvious that what has been evolving is the *psychological basis for the behaviour*, and this has evolved in quite different environmental conditions to the *environment of evolutionary adaptedeness*, or *EEA*.[90] Instead, they propose an alternative research programme.[91] The basic theses of the nativists are the following (Cosmides & Tooby

[90] Both the term and idea come from the British psychiatrist John Bowdy (1969), who explained some features of child development as being functional in the environment in which humans evolved (Laland & Brown 2002: 161.) Leda Cosmides and John Tooby (1987) adopted this notion and started to identify it roughly with the Pleistocene, during the last epoch in the history of our species in which the species was more or less one population and living in more or less similar conditions for long enough for evolutionary adaptation to have time to take place. The idea is contested: there are good reasons to think that the environment was not homogenous during this period (see Boyd & Silk 2003; Levin & Foley 2004), and human evolution did not happen exclusively during that period – the evolution of human psychology is in continuation with pre-human psychology, and there have been evolutionary changes after that period, too, even if the EEA was the most relevant period.

[91] This branch of evolutionary psychology can be argued to meet the criteria of a research programme *in sensu* Lakatos (1970) and should maybe evaluated as a research programme instead of as a series of empirical claims in its critical evaluation – but this is not important here.

1997, 2000, 2005a; Tooby & Cosmides 1989, 1990a, 1990b, 1992; Daly & Wilson 1999; Buss 1995 & 2014), my paraphrasing:

A) The mind has a *structure* that consists of *functional sub-parts* (*modules*) that perform specific tasks in cognition and motivation.

B) External behaviour varies for several reasons, including because of the different cultural surroundings people grow up in, but all the variation is dependent on the *species-typical* overall structure that is *innate*.[92]

C) The tasks that the mind is designed to perform are *domain-specific*, and they have been selectively evolved to perform these tasks in these domains to fulfil specific needs. In other words, they are *specific adaptations to specific environmental challenges*.

[92] The criticism of evolutionary psychology, in this narrow and a broader sense alike, that points out to variation across cultures, is somewhat misguided. Explaining cultural variation and underlying psychological structures are two separate projects that are not directly competitive (Mallon & Stich 2000). However, the range of actual variation and the degree to which the psychology is species-typical are empirical questions that matter to how relevant evolutionary psychological explanations are and how much understanding they can provide. The programmatic claim that evolutionary psychology reveals the universal human nature that should be the basis for all human science (for example, Tooby & Cosmides 1990b & 1992) is not very credible, for example (see Buller 2005; Richardson 2007; Ylikoski & Kokkonen 2009; and Smith 2020). If the universal features that evolutionary psychology reveals of human psychology were on the same level of abstraction as, say, the universality of language in the Chomskyan linguistics (in any of its forms; see Chomsky 1980, 1986, 1993 & 2000), this would be an interesting fact to learn about this capacity, but it would not help us to understand human behaviour any more than Chomskyan Universal Grammar would help us to understand any particular natural language.

D) The adaptation of the entire psychological architecture takes time. This is why the characteristics of mind that can be accounted for from the evolutionary perspective are rather species-typical than local, and also why the *Environment of Evolutionary Adaptiveness* (EEA) cannot be the current or a recent environment, but something that remained similar enough for a long enough time.

A few notes on these theses. The idea of modularity (A) and domain-specific adaptation (C) usually go hand in hand. A starting point for this is the standard approach of cognitive science that, following David Marr (1982), identifies three levels of analysis: the *computational* level (what task is accomplished), the *representational* level (how the system operates on the level of representational and motivational – that is, psychological – description), and the *implementation* level (how the brain operates). The evolutionary approach is a perspective on the computational level: what are the functions of mental capacities, cognitive mechanisms, emotions, and so on? Prima facie, this makes sense: if the operations of the brain and the resulting behaviour have evolutionary functions, they must be about what tasks are to be accomplished. This also provides an answer to what kind of functional design the psychologists and cognitive scientists should be looking for (if adaptive evolution is what explains this design), and how to justify the presupposition (or explain the fact) that cognitive processes have such designs in the first place. There are, however, four well known problems in this approach: 1) How modular is the mind, if at all[93]? 2)

[93] The concept of modularity in psychology comes from Jerry Fodor (1983). His list of criteria for modularity include domain specificity, automaticity, informational encapsulation, fastness of the processes, superficiality, specialization, and universal individual development. This is not an exhaustive list, but a list of some typical characteristics that distinguish some cognitive processes from creative, rational "central processing". Contemporary cognitive science distinguishes between two kinds of systems or domains of processes in a similar way without making references to modularity of the non-conscious

What is the right grain size – both for the tasks in the selective environments and the mental functions, they being bound together[94]? 3) The assumption of innateness and species-typicality of the mind's structure.[95] 4) The problems of adaptationism, which has been discussed already. Evolutionary psychology is explicitly making historical adaptation explanations, but this is not necessary. An *extended artefact model* version of evolutionary functionalism (including knowledge of actual evolutionary history) as discussed above, would probably be adequate to justify the research – although this interpretation also makes its claims to unificatory power (Tooby & Cosmides 1989 & 1992) much weaker. The first two problems, in turn, are empirical matters that affect *how* to do evolutionary psychology.

If the mind (or brain) did not possess evolved functional structures at all, this would be crucial for the existence of evolutionary psychology. If the mind is massively modular with precise tasks, studying its structure, both empirically and through evolutionary theorizing, would be much easier than if the structure is non-precise and the tasks are relatively general, and several behaviourally distinct traits

---

processes – I will return to this in the next chapter. In biology, there are other concepts of modularity, such as developmental modularity (Raff 1996), which may or may not be relevant to evolutionary psychology (see Wagner & Wagner 2003) but are not important to the discussion in this thesis.

[94] Anderson 2007 presents an idea of massive micro-modularity as a more realistic view of how mind works: the processes of mind (whether modular or not) use a set of sub-processes that, in turn, use sub-processes, until the "bottom level" of modular processes is reached. Non-modular processes can be constituted by modular processes. If this were the case, it would make more sense to think that the mental units that get selected are the constituent parts, on the basis of the multitude of tasks in which they participate, not task-level modules. This would make the search for modules as solutions to evolutionary challenges extremely complicated and difficult.

[95] I will discuss innateness later. The assumption of species-typicality is a more general problem for psychology: much of the experimental work has been done with a specific population and many of the results seem not to be cross-culturally generalizable (Henrich, Heine & Norenzayan 2010).

utilize the same structures.[96] So, for example, if there is a "cheater detection module" (Cosmides & Tooby 1992 & 2005b) that explains why the rule of material implication is automatically applied in the context of social norms but difficult to apply in some other contexts of inference (Wason 1966), the evolutionary explanation may be quite straightforward. If the explanation of this discrepancy in thinking comes from a more general point about relevance (Sperber *et al* 1995), the evolutionary connection between the environmental challenges and the way mind works becomes vaguer and more complicated – this holds even if the social norms constituted the actual selective context for the ability to use material implication in thinking. However, this does not mean that evolutionary psychology must be built on the idea of modularity, specific tasks, and specific challenges – just that it would be an easier discipline to practice if this were the case. A charitable interpretation of the nativist evolutionary psychology would be that it is the first step towards human evolutionary psychology[97] and just tries the easy way first. Bolhuis *et al* 2011, for example, calls for integration of criticism (about EEA, species-typicality, and massive modularity) into evolutionary psychological practices instead of considering them as a theoretical challenge to the discipline as such. Furthermore, the modularity approach may work for some but not all characteristics of cognition and motivation (see Atran 2005). This is an empirical issue not to be taken stance on here. It should be noted, however, that modularity is not necessarily connected to innateness:

[96] For arguments for massive modularity, see Sperber 1994 & 2001; for criticism, see Karmiloff-Smith 1996; Buller & Hardcastle 2000; Buller 2005; and Samules 2005.

[97] Evolutionary approaches can, naturally, be applied to animal cognition too. Interestingly, Steven Mithen (1995 & 1996) has argued that even relatively close ancestors of humans must have had modular minds, but the last stages of human evolution have been about an increase in the fluidity and sociality of mind. Even if this is the case, it is very unlikely that this would have wiped out all of the modular structure: evolution builds on existing structures rather than replacing them, here as well.

Annette Karmiloff-Smith (1996), for example, has argued that even modular structures of the adult mind are products of flexible individual development, not innate (see also Buller & Hardcastle 2000).

However, the grain size problem (and the identification of the evolving trait in general) has consequences for the issue of individualism and holism. Theoretically at least, the trait may be individual or interactional from the evolutionary functional perspective in two ways. First, there is the issue of how to define the phenotypic trait that is responsible for the fitness consequences. That is, *which set of capacities and resulting behaviour constitute the unit for the functionalist analysis*. Second, there is the issue of what the mechanisms of *inheritance* for this trait are. This is connected to thesis (B) above: the assumption of species-typicality and innateness, which usually go hand in hand.[98] The psychological disposition that is selected and inherited is, of course, a characteristic of an individual. The function it has (that is, the focus of the evolutionary analysis and explanation) is the role that this behaviour plays in the individual's life that affects the individual's fitness systematically. This may depend on the social surroundings in two different ways. Social evolution is a dynamic process where the interactions between the individuals and the behavioural dispositions (or the frequency of different dispositions across the population) depend on the makeup of the rest of the population and/or the group structure. In an individualistic approach, the rest of the population functions as a selective environment for the individual psychological traits. In an alternative interpretation, some behavioural *explananda* are *interactions*. The interactive phenotypes are functionally

[98] The target of evolutionary psychology is something that is thought to be species-typical (and therefore an object of species-wide evolutionary explanation), and the basis for species-typicality is the fixedness – that is, innateness – of the trait across the species. The contrast here is with individual or cultural traits that are acquired (not innate) and (therefore) not an object of species-wide explanations. However, the link between innateness and species-typicality does not need to be this strong. Furthermore, *both* concepts in this equation are vague and problematic. I will return to this later.

explainable as being more advantageous for the individuals who are participating in them, in contrast to the other alternative interactions, and this is not reducible to individual-level traits alone – this means that in the cases of this kind a holistic approach is needed.

4.2.2. Individualism and Holism in Evolutionary Psychology

Evolutionary psychology tends to be methodologically individualist in identifying the objects of research. Furthermore, evolutionary psychology tends to concentrate on traits that are innate rather than acquired. Even the nativist evolutionary psychologists do not dispute the role of culture and individual learning in psychology, but they claim that the proper target of evolutionary explanation is the psychological structure that is species-typical and innate and does nevertheless manifest even in behaviour affected by cultural and other acquired tendencies. The meaning (as well as the very meaningfulness) of "innateness" in these contexts is disputed. I will later articulate a concept of innateness that should both escape the standard criticism and capture the function the concept has in psychology. Under this interpretation, nativism would be about the identification of what is being explained, not a thesis about mind. If there is a psychological developmental process in which both environmental and genetic developmental factors contribute, both sets of factors constrain the possible outcomes. "Innate" in this context can be charitably understood as those aspects of psychology that are not affected by normal variation in environmental factors – that is, even if the environmental factors contribute to the developmental process, the result will be roughly the same under any combinations of causal factors present (within a reasonable degree of variation in these factors). If there is significant variation in the results of the developmental process that depends on the environmental factors, only those aspects of the trait that are not affected can be explained by evolution. This means that the *phenotypic plasticity* of a given *psychological* trait may have an evolutionary explanation, but not the details of the psychological

variants, nor the partly culturally transmitted *behavioural* traits that this evolved psychology underlies.

Once more, the individualistic approach is not the only possibility. The group/culture-dependent variation in psychological characteristics can also be an evolutionarily functional set of responses to changing environmental conditions in which the properties of the surrounding group in childhood serve as triggers or contributing factors for the development of the variant that is functional in the given context. This could be a case of adapted plasticity in the sense that environmental triggers choose between developmental pathways, directed learning, selected culture-dependency, or a combination of these things. Kim Sterelny (2003, 2007, 2012 & 2013), for example, has developed an alternative framework in which the changing social surroundings and active teaching play a role in evolutionary explanation. Karola Stotz (2014) has suggested a theoretical integration of the extended evolutionary synthesis (including non-genetic inheritance, plasticity, and other developmental aspects). The so-called "4e" model of cognition (embodied, enactive, embedded and extended; see Menary 2010) offers an alternative, extended approach to evolutionary psychology, challenging individualism on both cognition and its development.[99] Darcia Narvaez (2014; Narvaez *et al* 2012; Narvaez *et al* 2014) has applied evolutionary considerations to psychological *development* instead of the results of that development, tracing environmental (including social) difference-makers in development using evolutionary considerations. Christina Moya and Joseph Henrich (2016) have proposed a gene-culture co-evolutionary model as a perspective on psychology and the cultural variation in it, to replace nativist evolutionary psychology with *culture–gene coevolutionary psychology*. (See also Schaller *et al* 2010 for integrating evolutionary and

[99] The evolutionary psychologist Louise Barrett (2011) has also attempted to synthesize embodied cognition and cognitive ecology with traditional evolutionary psychology, although her approach is still individualistic.

traditional cultural psychology.) Evolutionary psychology does not need to be nativist nor individualist.

If plastic psychological traits are approached holistically in the developmental dimension, variation (and the range of traits), rather than merely species-typical universals, could be an object of evolutionary explanation. Variation can also be explained from an individualistic perspective: as mentioned, plasticity as such is easy enough to explain as an individual response to changing environments, and a certain range of variation or variant types, for example, can be explained through frequency-dependent selection of individual traits. But if there is functional variation between groups, in which the differences between groups are causal factors in individual development, an approach sensitive to group differences is needed, and individualist evolutionary psychology cannot explain the variation. Moving into explanation of culture-specific variants is moving from evolutionary psychology into evolutionary anthropology, but even some of the species-typical characteristics of the human mind may be products of this kind of evolution rather than individualist evolution, and need to be accounted for as such – biological evolution of the mind's capacities and culture's shaping of the mind should probably not be seen as temporarily consecutive stages but in interaction (see Tomasello 1999; Baumeister 2005; Schaller *et al* 2010; Sterelny 2012). Once again, it is a matter of where the empirical research leads, whether a holistic approach is ever needed, but there are methodological alternatives even within evolutionary psychology.

The final element in the list of the theses of evolutionary psychology is the assumption of the evolution of relevant traits taking time. This assumption fixes the objects of explanation as being innate and species-typical, and it is a distinguishing factor between evolutionary psychology and sociobiology (see Cosmides & Tooby 1987; Tooby & Cosmides 1989) as well as between evolutionary psychology and evolutionary anthropology (see Smith 2000; Laland & Brown 2002; Brown & Richerson 2013). This assumption may be considered part of the definition of what kind of evolutionary explanations, and therefore

what kind of explanatory targets, the different branches of evolutionary social sciences are interested in, more than a general ontological, empirical, or methodological assumption about evolution, natural selection, evolutionary explanation, or human psychology. However, this also means that the proper scope of evolutionary psychology, for example, may be significantly smaller than what evolutionary psychologists hope for.

In this section I have outlined some reasons to be interested in the difference between individualistic and holistic explanations within evolutionary psychology when it comes to the connections between evolutionary explanations and the proximate and developmental assumptions made by these explanations. In the previous section I provided some preliminary considerations for distinguishing between individualistic and holistic approaches in the evolutionary explanations proper, and these considerations hold for evolutionary psychology, too. Next, I will move on to evolutionary anthropology, in which the target of explanation is behaviour, not its psychological basis. Culture-dependent behavioural differences are a part of the explanatory agenda, and cultural inheritance, complex social traits, and the explicit use of group selection models make all these three issues more important.

4.3. Evolutionary Anthropology

Evolutionary anthropology, too, can be defined in more or less limited ways. Narrowly understood, it is the branch of biological anthropology that concentrates on the evolutionary history of human lineage, which includes paleoanthropology and comparative primatology. More broadly understood, it includes all evolutionary approaches to humans, including evolutionary functional considerations of human anatomy and physiology, as well as psychology, in which case it would include evolutionary psychology as well. In fact, biological anthropologists often include evolutionary approaches to mind in their toolkit (see, for example, Deacon 1997 and Fisher 2016). But since the

interest here lies in explaining behaviour, I will concentrate on the anthropologists' unique ways of explaining human behaviour with evolution directly, which includes research programmes such as behavioural ecology, theories of cultural evolution, dual inheritance theories, and the niche construction theory of culture. But since our interest lies in evolutionary explanation of human behaviour as a biological phenomenon, broadly speaking, I will not include evolutionary change within specific institutional settings.[100, 101]

Historically, evolutionary thinking was a part of anthropology in its "progressivist" interpretation – that is, evolution was understood

[100] "Special institutional settings" here would be things like a specific economic system such as free market capitalism (and the use of evolutionary game theory in understanding its dynamics; see Hargreaves Heap & Varoufakis 1995) or science as an institution (for evolutionary perspective to science, see Hull 1988b). Fundamental issues regarding these models are, however, presumably similar to what will be said about theories of cultural evolution in general.

[101] Another way to distinguish between evolutionary psychology and evolutionary anthropology would be via reference to the encompassing *institutional discipline* – that is, whether it is psychology or anthropology. There are, however, reasons to not to use this as a demarcation. First, there is a difference between doing epistemology-driven philosophy of science and institution-driven sociology of science, implying different criteria for distinctions. Second, the relevance of institutional divisions to the analytic understanding of epistemic issues related to disciplines as epistemic projects is questionable. Sometimes the epistemic connections between two subfields of different disciplines may be stronger than the connections between subfields within the same institutional discipline. From an epistemic point of view both psychology and anthropology are divided into differentiated epistemic projects that are more like different disciplines than organized sub-disciplinary parts of the same discipline. Third, the institutional location of various evolutionary approaches is not unambiguous, as can be discovered by a quick look at the variation in the academic backgrounds and the institutional settings of the research performed by the practitioners of these various branches of evolutionary human sciences. But mostly, the focus here is on the methodology, not institutions, and the classification adopted reflects this, not institutional structures.

as a progressive change towards more and more perfected forms, and different "cultures" (culturally different groups, typically identified with ethnic groups) were seen as more or less progressive in this sense. The various parts of anthropology studied different aspects of cultures that were seen as being at different stages of progress, and there were measures, for example, of different stages of linguistic evolution, the Western European languages being, naturally, on the top of the evolutionary scale. (Ingold 1995b.) When these ideas of progression were abandoned (rightly so, from the evolutionary point of view as well), most of anthropology resigned from evolutionary considerations altogether. Biological (or physical) anthropology remained the sphere for evolutionary considerations, and cultural anthropology concentrated on the differences between cultures that were explainable, basically, by their being of different cultures (Kuper 1999). Since the 1980s, however, evolutionary theory has made a comeback in the study of culture in three different ways: in *behavioural ecology*, in the dualistic *gene-culture co-evolution*, and what I will call the *evolution of cultural representations*, which includes several theories (e.g. *memetics* and *epidemiology of representations*).

4.3.1. Human Behavioural Ecology

Human behavioural ecology (for example, Chagnon & Irons 1979; Smith & Winterhalder 1992; Krebs et al 1997; Cronk et al 2000; Borgerhoff Mulder & Schact 2012) is a branch of anthropology that approaches human social behaviour (such as mating and marriage, parenting, and moral practices) as an adaptive reaction to local ecological conditions. It uses methods, concepts, and theories from evolutionary biology and sociobiology to do so, and it was partially an anthropological outgrowth of these (see Laland & Brown 2002; Brown & Richerson 2013). Unlike evolutionary psychology, it sees human behaviour as flexibly adaptable to different conditions, but unlike classical sociobiology, it does not bind this to genetic adaptation. One of the central theses of human behavioural ecology is the so-called *phenotypic gambit* (Grafen

1984): the phenotype (behaviour) is approached as adaptive to its current ecological setting, but the means by which it adapts is left black-boxed. It is assumed that this adaptation is a combination of evolved psychology (including evolved environment-sensitive decision rules both on proximate and developmental dimension), cultural adaptation, and individual intelligence (and the use of other general cognitive capacities) to make the best of the environment. The emphasis is on the environment (hence "ecology"). The measure of the success is, however, the fitness effects of the behaviour, not just fulfilling the psychological needs of the individual or the group, making this a form of evolutionary functionalism. Human behavioural ecology uses the form of evolutionary functionalism I labelled as *current utility analysis*, in contrast to the *artefact model* of which some of the evolutionary psychology is an example, and the *historical evolutionary functionalism* that characterizes some of the evolutionary psychology and that classical sociobiology got stuck with. The anthropologist and behavioural ecologist Monique Borgerhoff Mulder, for one, is quite straightforward on this point in constraining behavioural ecology to Tinbergen's question about functionality and leaving the other questions to other fields of study, including social sciences (Borgerhoff Mulder 1991: 69). However, anthropologists who use ecological tools do not restrict themselves to this, and it is not unusual for them to study the proximate mechanisms too (see Brown *et al* 2011). Much of what human behavioural ecologists do in practice is regular anthropological fieldwork. Human behavioural ecology is just an additional perspective that has become a school of its own.

4.3.2. Cultural Evolution

Social learning and culture are exceptionally important in human adaptation (Boyd, Richerson & Henrich 2011; Mathew & Perreault 2015). Behavioural ecologists grant culture a significant role in human adaptability, but they do not study the mechanisms of or the basis for cultural evolution. The concept of culture is itself a contested concept.

There is no unified view in cultural anthropology, for example, on how to define culture and what it is supposed to include: symbols, values, and representations only, or material products, technologies, and social structures as well? Is it something in the mind or in the external (but socially transmitted) behaviour? Is it a characteristic of an individual or a group? Are there holistic "cultures" or more atomistic "cultural features"? (Silverman 2002; see also Kuper 1999; Fox & King 2002b; Ramsey 2013b; Koskinen 2014.) Whatever it is, it is a non-genetic difference maker between different groups. For now, we can distinguish between two concepts of culture that refer to this: culture as a *medium* and culture as *content* on this medium. Culture as a medium is the biological phenomenon of humans having cultures. It includes whatever capacities are needed for this (social learning and language, for example) and the fact that humans transfer "cultural things" using this medium. Culture as content includes whatever "things" that are transferred. The biological question of whether a species of animal has a culture and if so, what is the nature of this culture (for example, Hunt & Gray 2003; Laland & Galef 2009; St Clair & Rutz 2013; Ramsey 2013b). The controversies about culture in cultural anthropology are about the nature of the contents of this medium.

Biological anthropologists and primatologists are often more interested in the medium of culture and how it functions and constitutes a non-genetic route for traits to spread in the population (for example, de Waal 2001; see also Ramsey 2013b), although sometimes the capacity to generate behavioural differences between groups is mentioned (McGrew 1998: 305). There is also a further issue about the *accumulation* of the modification (Boyd & Richerson 1985; McGrew 1998; Tomasello 1999; Carruthers 2013a) that may be required for genuine cultural evolution. The bioanthropological notion of a medium for sociocultural transmission is what is meant by culture in this dissertation (with some further distinctions later on), and as for what kind of things belong to it, there is no harm in being an inclusive pluralist here. The important thing for the main issue is that how individuals behave in social situations is affected by what they have learned from other

individuals, one way or another (for example, by copying behaviour, learning norms and attitudes, or developing meanings and beliefs that affect how the context is interpreted etc.). There are, however, theories about cultural change and cultural differences specifically.

One type of things that belong to culture are representations that are acquired from others and build our beliefs and affect our motives. There are several theories that try to capture the **evolution of cultural representations**. Edward O. Wilson presented an idea of "culture genes" as an elaboration of sociobiology (Lumdsen & Wilson 1981), and Richard Dawkins (1982) presented the similar idea of *memes*, later developed into *memetics* by Susan Blackmore (1999). The general idea in these theories is that there is something analogical to genes that copy themselves from one mind to another, and there could be an evolutionary theory of cultural change. It is, however, difficult to capture the nature of the entity that is supposed to be the self-replicating unit that forms lineages (see Boyd & Richerson 2000; Sperber 2000; Sterelny 2006a; Lewens 2015). The lack of mechanistic detail in the theory leaves memetics explanatorily empty and merely an *alternative narrative* for cultural change (see Lewens 2015). A more sophisticated version of the same idea is Dan Sperber's (1996) theory of the *epidemiology of representations*, which pays attention to culture-related cognition. Its main idea is that ideas, norms, et cetera are copied from human mind to human mind through communication, in which both the process of communication and interpretation bias what is transmitted, as do the characteristics of human cognitive architecture, such as what concepts, narratives et cetera are easy for us to understand and remember. Human mind is the environment to which ideas must adapt, and the features of the human mind can explain, for example, what religious ideas are likely to emerge (Boyer 2001; Atran 2002). The epidemiology of representations is partly based on the evolutionary psychological understanding of mind, although what matters for it is only the understanding of how mind works, not its evolution, and the researchers of the field are usually more interested in cross-cultural comparative psychology (or cognitive anthropology) than evolutionary psychologists

proper who are more content with monocultural laboratory studies. None of the evolutionary theories of cultural representations specify the connection between genetic and cultural evolution; they are theories of cultural evolution only. Consequently, the relevance of these fields for the issue at hand is limited.

4.3.3. Genes and Culture in Interaction

Culture, as the medium, is a more significant issue. Culture in this sense is a product of evolution itself. In some cases, selection favours automatic reactions and fixedness of the behaviour, while in other cases, it favours developmental plasticity (West-Eberhard 2003). The forms of plasticity range from the triggering of a fixed and adapted, alternative developmental pathway by environmental conditions, to learning something new through individual creativity and intelligence. As a rule, learning is biased, directed, and constrained to some degree. The advantage of learning is the potential for novel behavioural responses during an individual's lifetime, while genetic evolution takes generations. Genetic evolution, however, selects behavioural responses according to their actual fitness benefits without the need for an individual considering, which has a higher risk for error, which makes genetic adaptation better than learning in most cases, if there is time to adapt. Individual learning and intelligence are favoured in conditions where constant and rapid adaptation to new circumstances is needed. (Sterelny 2003.) Culture and social learning make behavioural change faster than genetic evolution and slower than individual learning; it allows cumulative learning that surpasses everything that an individual could accomplish, but it is also vulnerable to similar errors as individual learning – there is more time to correct, but the possibility of multigenerational accumulation of errors, too (Boyd *et al* 2011). Social learning and cumulative culture are evolutionarily useful only under quite specific context (for example, when a species is spread over a vast range of ecological conditions that change over time but stay the same for several generations, and there

is gene flow between the groups). This is why social learning is rare in nature.[102] (Boyd & Richerson 1985 & 2005; Boyd & Silk 2003.)

The specific contents of cultural representations also influence behaviour in ways that have fitness consequences. If the capacity for culture is evolved, it is to be expected that humans rely on social learning and individual thinking to different degrees in different contexts, and what and how is learned is constrained in ways that make evolutionary sense. Culturally invented and acquired capacities and habits may also become increasingly independent on external information in their development, if there is a selection pressure to acquire them effectively, through the *Baldwin effect*.[103] All this means that the role culture has played in human evolution should inform evolutionary approaches of psychology, not only the other way around. But even on a shorter time scale, culturally acquired behaviour may have fitness consequences.

The theory of **gene-culture co-evolution** (or *dual inheritance*) (Boyd & Richerson 1985 & 2005; see also Durham 1995) explains human behaviour as a product of interconnected genetic and cultural evolution. For the purposes of the theory, "culture" is defined as "information capable of affecting individuals' behaviour that they acquire from members of their species through teaching, imitation, and

[102] New Caledonian crows seem to be the sole example of cumulative culture outside the human species (Hunt & Gray 2003; St. Clair & Rutz 2013). The lack of social learning is likely to be a matter of willingness to do so rather than the lack of the necessary cognitive capacities – this seems to be the case of the famous kea parrots, at least, who are able to learn by observing others but seem not to do so in nature (Gajdon et al 2004) – and there may be an evolutionary reason for this preference.

[103] If a characteristic that is acquired is useful, more and more powerful ways to acquire it get selected, and genes that guide the development of the characteristic may be selected, too, if they emerge (Baldwin 1896; Futuyma 1998). The same applies in the other direction as well: if a developmental resource is reliably available in the environment, genetic guidance of the developmental process may disappear.

other forms of social transmission", where information means "any kind of mental state, conscious or not" (Richerson & Boyd 2005: 5). Social learning involving information transference is biased by both general and individual capacities and biases in learning, memory, and heuristics about who to copy from, but its assumptions about human psychology involve domain-general rather than domain-specific capacities, in contrast to nativist evolutionary psychology (Sterelny 2012; Brown & Richerson 2013; Carruthers 2013a). The heuristics are biased in, for example, which individuals are more likely to be copied, based on their social status, perceived success, *et cetera*. These biases and other contextual differences in when to copy others are products of biological evolution: we have the tendency for social learning (and teaching) in the contexts where it has been evolutionarily more beneficial to rely on other than rely on one's own thinking or (adapted) inclinations. This means that there is an evolutionary loop between cultural and genetic evolution: local cultures are evolving based on the biological success they bring, and the biological evolution of psychology is adapting to changing cultural conditions and the needs to rely on culture and social learning. Both cultural and genetic evolution are on-going processes, although cultural evolution is faster, as well as more directed, since new variation (produced by individual innovation) is directed, and transmission is biased (according to the behavioural trait's content and context). However, if the dual inheritance theory has an approximately correct image of human evolution, it is important to understand that whatever species-typicality there is in human psychology, it is partly adapted to a socio-cultural environment like this, and to an on-going cultural evolution and its needs.

The dual inheritance theory has three key features that make it a holistic approach. First, the cultural transmission is horizontal, harmonizing the behaviour on the group level, and relies on collective properties of the group (for example, the biases in who to copy from depend on the relative success and social status relative to other individuals). Second, there is an idea of *cumulative culture*: behaviour, technology, *et cetera* become more adapted, more sophisticated, and more

complex from generation to generation, surpassing individual limitations. This may lead into behavioural traits that are increasingly dependent on the group level structures and interactions with others that make them holistic in the sense that I will discuss in the next two chapters in more detail. Third, the key idea in connecting culture and genes is *cultural group selection*. Despite the name, this involves not only selection between cultural groups (as cultural quasi-organisms), but it is also a form of biological (gene-based) group selection between groups that have different behavioural traits based on cultural differences through the co-evolutionary loops described above.

4.3.4. Schools in Comparison

I will not discuss any of the "schools" of evolutionary human social sciences any further, nor will I try to evaluate them or discuss their compatibility. Any of them may be an adequate explanatory perspective on some human behaviour[104], and much of human behaviour is probably such that the evolutionary perspective has nothing, or only trivial things, to say about it – but all this is a matter of empirical study and trial of various explanatory strategies. But there are some methodological issues that have come up and will be the topics of the following chapters. First, the role played by evolutionary functionalist thinking is different in different schools, not just what they say about behaviour or its evolution. Classical sociobiology was a historical adaptationist project, and so is co-evolution theory. Some forms of evolutionary psychology also fall in this category, but some of it can be thought of as an artefact model instead, while behavioural ecology analyses behaviour from the current use perspective. Secondly, the very targets of the various evolutionary approaches seem to be different. Evolutionary anthropology is interested in the behaviour while

[104] Smith *et al* 2000, for example, identify contradictions between evolutionary psychology and human behavioural ecology while advocating pluralism of approaches.

evolutionary psychology is interested in the psychology that produces it (although some anthropologists are interested in both). As I pointed out earlier, it is not necessary for these to map onto each other: a psychological mechanism implies a set of behavioural dispositions, but the functionally adequate partitioning of behaviour into behavioural traits may be different, for both ecological and co-evolutionary explanatory purposes.

Different uses of evolutionary functionalist thinking or different *explananda* do not, however, make these approaches different accounts of what evolutionary human behavioural science should be all about – they have different *explananda*. They are not in direct competition, but neither are they unconnected projects. The assumptions they make may be mutually incompatible, such as about how mind works and what the role of culture in the individual development is. I will analyse the relationship of the different *explananda* in general in the following chapter, and with the example of cooperative and altruistic behaviour in chapter after that. Cooperation and altruism have been central to all approaches (see Brown & Richerson 2013) as well as being key topics of discussion in the philosophy of biology. I will argue that sometimes the *explanandum* of evolutionary anthropology is a type of interaction that cannot be understood from an individualist perspective, but only as an interactive trait. But I will even go a step further: since the evolutionary perspective of the evolutionary psychology has to do with what task the psychological characteristics are selected for, and in the case of social behaviour, this is participation in the selected interaction, evolutionary psychology cannot be individualistic in all these cases either.

Another issue has to do with development. Evolutionary psychologists tend to be interested in innate traits, whereas anthropologists are also interested in the behaviour that is influenced by culture. Regarding this, I will discuss the concept of innateness and defend a way to understand psychological innateness that makes sense in the context of evolutionary psychology. I will also discuss culture as a holistic route of inheritance for reproducing behavioural traits in the

next generation and what difference it makes. The third and final issue, and the main issue of the dissertation, is group selection. Co-evolutionary theorists fully embrace this as an explanatory tool, in contrast to evolutionary psychologists who tend to be individualists in this dimension as well. Behavioural ecologists are a more difficult case: they approach issues individualistically while acknowledging culture as a factor. My main contribution to the philosophical discussion on group selection will be to show how the previous two issues are related to the group selection controversy in the human context, and how this helps us understand the contested issue of what counts as group selection in the first place.

5. On Human Behaviour and Its Causes

Dividing behaviour into explainable biological traits is more difficult than doing so in physiology, or even in psychology. There is no unambiguous way to slice behaviour into behavioural traits in biology in general (see Levitis, Lidicker & Freund 2009), and it is even more difficult to do with human behaviour (see Longino 2013). In general, behaviour can be defined as the activity of a system, such as the bodily movements of an animal or a human being (Dretske 1988). In philosophical psychology and action theory, behaviour as a mere bodily movement is usually a vague notion merely to be contrasted with *action* as something intentional that the agent does. This contrast involves attributing the agent reasons that identify the action. From this perspective, behaviour as bodily movement is understood as a component of action, and there is no need to scrutinize behaviour independently of this. There is no need for an analysis of what counts as a behavioural trait; giving reasons (or *rationalizing* the action) provides all the structure we need to slice behaviour into the units we need, for the purposes of understanding human action in this way.

However, if we want to talk about behavioural *traits* as an object of explanation, for example, we need criteria for what goes together whether we use rational action explanations in understanding specific instances of behaviour or not. Furthermore, what counts as a trait depends on our purposes for partitioning behaviour into traits, including our current explanatory interests (see Longino 2013). Taking rationalization as a basis may serve this purpose in some cases, but it may be inadequate or even dangerously misleading in others. The latter is the case with evolutionary explanations: evolutionary psychology, for example, does not explain our reasons for action. I will comment on this in greater detail later. Furthermore, rationalization of action is intimately connected to individual psychology: what we understand the behaviour to be about is conceptually connected to the aims and beliefs of an agent and to what the behaviour means for them. Folk psychology does not make the clear distinction between

thinking and behaving in action that we need to make to separate behaviour and its underlying psychology as *explananda*. Yet, at the same time, folk psychology is a constitutive part of the human social practices, and folk-psychological attributions of intentional agency cannot be disregarded in the description of the *explananda*, even if folk psychology does not provide an adequate framework for conceptualizing the *explananda* on psychological or behavioural level, or for explanations. This is why I will use some space to comment on what folk psychology is from the evolutionary perspective, and how it is related to cognition, motivation, and behaviour as objects of scientific research, instead of just bypassing the whole topic, which is usually done in discussing evolutionary explanations of behaviour. Instead, I will make a distinction between psychological and agentive levels of description, which, I will argue, is not explicit within the folk-psychological attributions, and I will use this distinction in the next chapter in discussing social interaction.

The main topic of this chapter is to develop an evolutionary functionalist account for defining human behavioural traits and shows how they differ from psychological traits (as well as from agentive description) and how they are related. This is done in part by contrasting it to the folk psychological understanding of how psychology and behaviour are related and showing how folk psychology is inadequate. This prepares for the following chapter, in which I argue that behavioural traits, understood this way, can be supra-individual traits. I will also briefly discuss the role folk psychology and rationality assumptions play in the evolution of human social behaviour. I will start by setting the question, then discussing the nature of folk psychology and action explanations, and finally returning to the question of how to identify behavioural traits as an object of explanation.

5.1. Preliminary Issues

The standard way to approach social interaction from the evolutionary perspective is to understand behaviour as the interaction of individually implemented strategies, where the social surroundings are the selective environment only. The evolution of strategies is dynamic, with a possibility of large-scale emergent consequences, but the selection in this picture is between traits of individuals. The *sociality* of behaviour simply means that the behaviour has fitness implications for both the individual who performs the behaviour and some other individual(s) with whom the performing individual interacts. In other words, the approach is individualistic (when it comes to the proximate dimension). In addition, the behavioural dispositions are treated as isolated traits and the distinction between a behavioural trait and the mechanism producing it, which is left black-boxed, is not assumed to be significant. However, as we have seen in the previous chapters, since the same psychological faculties may participate in a range of different behaviours, the fitness of the faculties depends on the consequences of the resulting behaviour, and the resulting behaviour depends on the context as well, we need to keep psychology and behaviour separate even within the individualist context. This opens the door to considering alternative, non-individualist ways to identify behavioural traits. I suggested two alternative versions of holism in the introduction: *collectivist holism* and *interactionist holism*. In collectivist holism, behavioural traits are traits of a collective. In a human context, this would involve distinguishable groups that act as *collective agents* with respect to the behavioural trait in question. *Superorganisms* such as social insect colonies are groups that behave only in this manner, as if an individual (see Haber 2013). Some social institutions with intentional planning of their structure may qualify as examples (see List & Pettit 2011, for a candidate), but it is doubtful that natural human groups have evolved to possess such uniformity (although see Pagel 2012 as an example of a biological superorganism view of human groups). Shared cultural meanings and social norms can bring such

uniformity to groups regarding some behaviour, which makes this form of holism still relevant to consider with some individual behavioural traits. In the interactionist holism, the traits are forms of interaction between two or more individuals. The traits are not attributed to a collective of individuals or an individual but to sets of individuals whose behaviour in a context constitutes a trait that binds the individuals together in an *interactive trait*.

Superorganisms are quasi-individuals and automatically *foci* for fitness considerations as such.[105] I will discuss the issue of whether interactionist holism involves fitness considerations above the individual level as well (through *trait group* selection; Wilson & Sober 1994; Sober & Wilson 1998) in the last part of the dissertation. Individualism and holism regarding the behaviour and its mechanistic basis is a fundamental methodological issue in the evolutionary explanation of social behaviour in either case (see Moore *et al* 1997; Formica *et al* 2017). I will later argue that even in some cases where the evolution is modelled as individual action (for example, with game theory), where the models correctly show the selection *of* individual phenotypes, the traits that are selected *for* are the emerging interactions. Individual behavioural dispositions are selected by virtue of participation in these interactions. Before this can be argued, we must establish that such traits are possible in the human context.

5.1.1. Evolutionary Requirements for Interactive Traits

A central issue in what can count as a trait from the evolutionary perspective is *evolvability* (Wagner 2005; Pigliucci 2008). Evolvability requires heritability and a high enough degree of modularity for different traits to have separate evolutionary trajectories. For behavioural

[105] This is partly a definitional issue about what counts as an individual (see Bouchard & Huneman 2013). Charles Goodnight (2013), for example, defines an individual as the lowest level of biological organization that is relevant for (multilevel) selection models.

traits, this requires repeatability (see Formica *et al* 2017). These requirements are not a problem for interactive phenotypes if they are not a problem for individualistic behavioural phenotypes either. Social behavioural patterns are dependent on individual behavioural dispositions, whether we understand them as consequences of individual behavioural dispositions in social contexts or as higher-level relational traits. If there are evolvable individual behavioural traits, there can be interactionist traits constituted by such traits in repeatable interactions. Evolvability of behavioural traits as such does not seem to be a problem with repetitive behaviour (see Katz 2011) and repeatability may be very robust in higher-level social behavioural traits too (Formica *et al* 2017). Behavioural plasticity limits the level of detail in which traits can be evolvable, but cultural inheritance and social norms increase the robustness of social traits in more detail again. I will later effectively argue for the evolvability of interactionist and social-environment sensitive traits in humans, but I will not discuss evolvability itself directly. I will assume that some social behavioural traits are evolvable. The constraints posed by what is evolvable and what is not limit the scope of evolutionary explanations of social behaviour and therefore the scope of the subject matter of the discussion. None of this has any effect on the issue of individualism and holism within the scope of evolutionary explanations as such.

A further *prima facie* challenge to the idea of interactive traits is that they are both evolving traits and the environment for their own evolution at the same time. This is a bit odd. The evolution of social behaviour (social selection) is always a dynamic interplay between individual traits, but the individualistic models separate the evolving trait and its environment as a snapshot of the evolution of the entire population. The individual traits that constitute the interaction are co-evolving, interacting phenotypes that are each other's selective environment. However, the co-evolution of interacting phenotypes like this can be modelled using indirect fitness effects like those in kin selection models as seemingly individual selection (Moore *et al* 1997). The trait group approach (Wilson & Sober 1994; Sober & Wilson 1998)

is an evolutionarily holistic alternative to model some of the same interactions, although this model presents the traits themselves just as individualistically. But both approaches explain the form of interaction (interactive traits) while at the same time these interactions play a causal role in the explanation of why individuals with certain behavioural dispositions (individual traits) have higher fitness than some other individual phenotypes. This means that there are two phenotypes in the *explanandum*. To understand the causal structure of the selection process (and not just model it), we should distinguish between two different but intertwined *explananda*, rather than conflating them: the *form of behaviour* and the *proximate psychological mechanisms* underlying the behaviour. The proposition that behaviour and its proximate mechanisms are different objects of explanation may sound trivial but also non-consequential because of the intimate connection between the two. This may be the case with most behaviour, where the form of behaviour is simply a result of the underlying psychology and the context that triggers the behaviour. However, the very idea of interactive phenotypes is that the form of behaviour is interactive. This gives the difference explanatory significance. The forms of social interaction compete with each other while the individual behavioural dispositions that are the constituents of interactive phenotypes compete with other individual dispositions in being able to participate in these interactions. We need to understand both parts and this implies evolving traits on two different levels: individual and group.

5.1.2. Human Behavioural Traits and the Problems with Folk Psychology

The problem with distinguishing a form of behaviour and the psychological mechanism producing it is that we need criteria for what is a behavioural trait. What counts as "doing the same"? This is also important for repeatability and therefore for evolvability as well. One option is to describe the external physical behaviour as robust,

repeated structural patterns and simply lump all the behaviour with the same pattern into the same trait. The problem is that the same pattern of behaviour may have completely different triggering causes and consequences in two different contexts, and two different sets of bodily movements may have the same function in achieving something. For instance, swinging a hand in the air may constitute a greeting, but it may be an act of stopping a taxi or scaring a wasp away, and greeting can take place in other ways. Mere bodily movement out of context is not what we are usually interested in when explaining behaviour or how we understand behaviour in the first place.[106] We are not interested in explaining hand waving behaviour, but we might be interested in explaining greeting behaviour or reactions to intrusive wasps. The identity of the behaviour depends on the context too – it is both a reaction to something and directed to achieve something. How we characterize behavioural traits should reflect this, both in biology and psychology, whether or not we use folk psychology. (See Enc 1995.) Having a behavioural trait as an *explanandum* requires generalizability of the *kind of reaction* to a *kind of context* with an idea of what the behaviour *achieves* (or attempts to achieve) in the context – in other words, what it is *about*.

With simple stimulus-response systems, this would be a straightforward pairing of a behavioural pattern with an environment, but as the behavioural responses become more flexible, and the relevant context of behaviour becomes more and more inclusive of factors not concretely present, the identification of behaviour with environment becomes more complicated as well. Most human behaviour is related to goals and representations of things and states of affairs that are not visible in the external behaviour and its immediate environment alone. This alone does not mean that there are no behavioural traits or that they are not separable from each other. This only means that a trait as a pattern of movements and a trait as a functional reaction are

[106] Of course, it could be, if we are interested in neuromotor control of a bodily movement itself, for example, but this is not what we are usually after.

separate, and a heterogenous cluster of possible patterns of behaviour can constitute the latter. This makes it more adequate to give behaviour semantically loaded descriptions that refer to the goals and representations involved in the behaviour. Both everyday practices and many scientific theories understand human behaviour as goal-directed intentional action like this. This makes sense on the first approximation and makes a good candidate for what instances of behaviour should be categorized as "doing the same": the instances of behaviour that are executed with the same intention of achievement. Fred Dretske (1988), for example, suggests that reasons (or rationalizations) structure the behaviour in an adequate way by stating what it is *for*, which enables the identification of behaviour. Reasons are constituted by propositional attitudes such as beliefs and desires. Both evolutionary psychologists and anthropologists characterize human behaviour using folk-psychological conceptualizations like this. However, this can be misleading, for two reasons.

First, folk psychology is ambiguous as to whether (or rather: when) its concepts refer to personal and sub-personal levels. Consider the following example. A person, let us call him Amos, is drowning. Another person, let us call her Beatrice, sees this. Beatrice jumps into the water and rescues Amos. Let us suppose Beatrice has no other external goals for the action than to save Amos – she is not motivated by appearing heroic, getting gratitude, promoting romantic feelings, signalling her virtues, or any other purpose that she might use the rescue as an instrument for. We would say that the *reason* for Beatrice's action is her *desire to help* him and her *belief* that she can do so by jumping into the water and pulling him to shore. Furthermore, even if there are other reasons for her to jump into the water, or further reasons to save Amos, if this is the *primary reason*[107], we would say that this was her *intention to act*. Now, let us suppose that an egoistic motivation theory

[107] Primary reason is the "actual" reason, the one that made the person do what they did (see Davidson 1963; O'Brien 2019). I will return to the role of reasons and primary reasons in explanation later.

of helping behaviour is true (see Batson 2011). For example, the internal motivation structure is such that the causal process of perceiving Amos's distress and jumping to help him goes through the distress that Beatrice feels when she sees Amos in danger. This contrasts with an altruistic theory of motivation (or *empathetic concern theory*), according to which seeing *Amos's* distress is the motivating (or triggering) factor for Beatrice, and her own distress is not needed (Batson 1991 & 2011). In this case, the "ultimate desire" in the succession of desires giving rise to each other is an egoistic one (see Sober 1992c; Sober & Wilson 1998).

Both descriptions (that Beatrice's action is guided by her non-instrumental desire to help Amos, and that Beatrice's action is guided by her need to get rid of her own distress) seem to be right. Are folk-psychological attributions about the agent's overall goals or about the underlying psychological states? The debate over psychological egoism and altruism revolves around the issue of what the ultimate desires guiding the action are (see Kitcher 1997; Sober & Wilson 1998; Stich 2007; Batson 2011), but it is somewhat unclear whether these desires are meant to be attributes of the whole person (making Beatrice an altruist) or sub-personal psychological factors (making her an egoist). What does it mean to say that Beatrice is ultimately acting upon selfish motives, for example? The debate seems to assume that the desires of a person are attributes of the whole person and effective psychological states at the same time. I will analyse this in greater detail later. My suggestion is that there are two levels of description, **psychological** and **agentive**, and folk psychology does not make this distinction.[108] This in turn is important for how we think about the

[108] In social psychology, the object of study includes *cognitive*, *affective*, and *behavioural* aspects of sociality-related psychology. The "psychological" under my definition includes both cognitive and affective parts of psychology, but not behaviour. "Psychology" is all those processes that produce behaviour. In philosophy and behavioural sciences, it has become commonplace to concentrate on the relationship between cognition and behaviour. In some cases, "cognitive" would be practically synonymous with "psychological" and this

identification of behaviour and what its causes are, which in turn is important to knowing what it is that we are explaining in the evolutionary explanations of behaviour.

The distinction between propositional attitudes ascribable to a person as a whole and representational states participating in cognitive processes is not new, but much of the philosophical discussion on the matter has been about what is the *right* level of description. There are several theories of propositional attitudes and the nature of folk psychology, and for current purposes I will group them into *cognitivists,* who think that propositional attitudes are causally effective psychological states, and *ascriptionists,* who think that propositional attitudes are the states that we attribute to agents. Cognitivists are usually *representationalists,* who think that propositional attitudes are psychological entities, *representations*, which participate in cognitive processes (for example, Fodor 1975, 1981, 1990; Millikan 1984; Dretske 1988; Cummins 1996; Shea 2018), and that folk-psychological practices trace them. I will include *eliminativists* (P.M. Churchland 1981 & 1989; Stich 1983; P.S. Churchland 1986; Churchland & Churchland 1998) in this camp too, since what they are eliminativists about is propositional attitudes as participating in cognitive processes rather than agentive ascriptions. The ascriptionists include *interpretationists* (for example,

---

is the case in some discussions that I refer to. However, since a large part of the discussion will be about motivational mechanisms such as psychological egoism or empathetic concern, which are affective rather than cognitive factors, I will use the term "psychological" as the general term to capture all that is necessary. Furthermore, although I make the distinction between psychological and agentive that both are sometimes called "psychological" or "mental", and I will use finer distinctions between levels of abstraction in referring to psychological processes and mechanisms, such as Marr's levels (Marr 1982), these finer distinctions are all within what I call psychological. These terminological discrepancies reflect different conceptualizations of the same processes, sometimes with different levels of abstractions as their explanatory perspectives – but the proposition I am about to argue for is that psychological and agentive are different categories altogether.

Davidson 1963; Anscombe 1967; Dennett 1971 & 1987) and *dispositionalists* (for example, Ryle 1949; Marcus 1990), who think that propositional attitudes get their semantics from patterns of action and reaction and that propositional attitudes are attributed to agents based on their dispositions to behave according to these patterns. Additionally, *functionalists* think that the best way to characterize propositional attitudes (or mental states in general) is through the role a state plays in relation to something else (other such states, perceptual input, behaviour etc.), regardless of which camp the person belongs to otherwise (for example, Fodor 1968; Lewis 1972; Putnam 1975; Dennett 1978; Millikan 1984; Shea 2018).

What I suggest instead is that folk psychology *as a natural practice* (and the basis for the conflicting philosophical intuitions regarding mental states) does not make this distinction in the first place. Its concepts may refer to holistic states of an agent (hence *agentive* level) or entities within the cognition and internal factors in action motivation (*psychological* level), depending on the situation, and the practice itself does not always recognize a difference between the two cases. Propositional attitudes understood as agentive states are intentional states attributed to an agent. Psychological states are causal factors within the functioning of mind, having to do with representations and motivational forces (or drives) that participate in cognitive and behaviour-guiding processes. The debate on the right level of description to think about the reference of folk-psychological concepts is also a debate on how to understand intentionality. This is one reason why I use the term "agentive" instead of "intentional." There may actually be good reasons for not distinguishing between the two levels in everyday practice, but this is still problematic in philosophy and in science. Our intuitions about mind and behaviour are based on our *use* of folk-psychological capacities, and its conceptualizations inform our understanding of human behaviour outside what is observed or explicitly hypothesized in any given theory. In other words, folk psychology brings in "meta-theory-ladenness" to both empirical and

philosophical research.[109] I will spend a big part of this chapter on this issue, for although folk psychology is not a part of the main topic of the dissertation, and I might be able to simply stipulate the distinction for what is required for the overall argument, folk psychology confuses intuitions about the subject matter (including in science), and folk psychology as a practice is a part of the explanandum of the evolutionary explanations under discussion (which is also the reason why our intuitions, as participators in this practice, are influenced by it). I will discuss some aspects of folk psychology both to convince that another approach is needed and to clarify what I am proposing. I will try to be as brief as possible, which inevitably leads to some superficiality considering the significance of the issue and the amount of literature on each detail, but my aim is only to show that the distinction I propose is needed and that it does not require whimsical revisionism. I will discuss both philosophical debates and the empirical research on folk psychology to argue for the plausibility of the suggestion.

The central issue is that the ambiguity of the reference of folk-psychological attributions fixes the focus on the individual by connecting the form of behaviour and the individual psychological dispositions directly. Intentional action descriptions identify both what the action is *about* (which we need in order to talk about behavioural traits) and what kind of psychological states are causing it. We categorize human behaviour according to what it is *intended* to be about,

[109] Consequently, if this position is correct, both empirical psychology (and philosophy of psychology as well as those parts of philosophy of mind that are concerned with the workings of mind) and action theory (and philosophical psychology insofar as it analyses personal-level states) must be revisionist, although in different "directions." The nature of these revisions is not a matter for this thesis, but I would argue that some of these revisions, or their essential features, have in fact been made on these fields already, even if they have been left unarticulated and some conceptual confusions remain. This in turn is the source of some apparent discrepancies between psychology and action theory. I will return to this to some extent later.

and this is inherently connected to intentional *psychology*.[110] Folk psychology seems to presuppose individualism and conflates the two senses of trait that need to be distinguished for the analysis of the evolutionary *explanandum* in social behaviour. This confusion can manifest in two connected ways: first, if we use a folk-psychological framework in describing psychological and behavioural traits (that is, we have folk-psychological assumptions within the substance of the research), and second, if folk-psychological presuppositions bias researchers' observations and categorization of behaviour itself (that is, researches apply folk psychology in making observations). For example, evolutionary psychologists explicitly explain human wants.[111] Thinking about behaviour in a folk-psychological framework may smuggle folk-psychological elements into both scientific observation and philosophical intuitions without epistemic justification, and this includes individualistic thinking about behaviour.

Another problem in understanding behaviour through folk-psychological conceptualizations is that folk psychology is primarily a framework to make sense of a *particular* individual action and is not sufficient to describe generalizable *traits* in the first place. Traits are behavioural tendencies rather than actions that are rationalizable in context. The object of an evolutionary explanation may be a psychological disposition or a behavioural pattern, but never an individual action. We can come up with a reason or a cluster of reasons that unify a series of actions, but this is different from a tendency – it is giving a reason for (or rationalizing) a tendency. We may do so and call this a

---

[110] The Logical Connection argument (von Wright 1971) states that since we attribute intentional states to agents based on our interpretation of their behaviour as intentional action, the intentional states cannot be psychological factors that *cause* the behaviour. What I am saying here is a weaker point about the connection: folk-psychological attributions identify action by intentions regardless of whether they are ascriptions or causal states, and this attribution fixes an isomorphic relation between agentive and psychological level structure.

[111] I will return to this "Freudo-Darwinian Fallacy" (Ylikoski & Kokkonen 2009; see also Buller 1999) later.

habit, for example, but a psychological explanation would rather refer to the tendency to have certain kinds of inclinations and reactions (or, in a rationalizing language, to come up with certain kinds of reasons instead of other possible reasons). Evolutionary explanations are about these features of the psychological makeup, not about reasons. This also means that if we distinguish between behavioural traits and psychological traits (as I will), agentive-level descriptions are different from both sub-personal psychological descriptions and behavioural descriptions. I will discuss this in greater detail later. In any case, behavioural traits should be objects of *generic explanation* (in whichever explanatory dimension) instead of a *contextual particular explanation* (which is the real focus of folk-psychological descriptions even if we use the same reason to explain multiple actions). This means that whether we use a folk-psychological framework or not, we still need further criteria for what behaviour goes together as a trait.

5.1.3. Non-folksy Alternatives

An alternative solution to the problem of defining a behavioural trait (without using folk psychology) would be to organize behaviour according to the psychological mechanisms that produce it. This might seem to be a sensible solution, especially in evolutionary explanations, given that psychological mechanisms are the traits that are inherited. This is the intuition that guides evolutionary psychology, as we have seen. There are problems with this alternative, too. For one, it is not clear what these mechanisms are supposed to be. As discussed in the previous chapter, this would probably work with the isolated innate modules approach of nativist evolutionary psychology, but this approach is probably wrong. If cognitive processes and mechanisms have a more complicated architecture, identification of psychological mechanisms turns out to be identification of capacities that are identified by their functions, and the functions are identified by tasks and other performance-related categories. Psychological characteristics are patterns discovered in research and the mechanisms that are

postulated to explain them. Furthermore, behaviour takes place in a complex network of causal factors (including various psychological processes and environmental factors) and the explanatory interests determine how to break down this network. Psychology is not the only sensible explanatory interest. For example, if we are interested in the role of behaviour in a social context, we might want to organize behaviour into traits in a different way. Furthermore, since the evolutionary benefits of psychological characteristics depend on the behaviour they contribute to in a variety of environments, we need to sort the behaviour into types of interaction anyway, even if we were interested only in the adaptivity of a psychological characteristic.

We could also slice the behaviour according to what factors *trigger* it. For any behaviour, we can distinguish between *triggering* and *structuring* causes (Dretske 1988): the structuring causes are the background conditions and mechanistic details that are presupposed, and the triggering cause is what the behaviour under scrutiny is a reaction to, internal or external to the system itself. In this image, triggers would identify the behaviour, whereas some of the structuring causes constitute a psychological mechanism, which is the psychological trait that explains the behaviour, other factors being background conditions. Again, how we categorize any particular factor (including what is included in the trait) depends on our explanatory interests. This causes the different approaches of different behavioural sciences, for example, with different explanatory resources, to have, strictly speaking, somewhat different *explananda*, even if we start with the same instance of behaviour.

As Helen Longino (2013) has pointed out, different human behavioural sciences define a behavioural trait using incommensurable categories: sometimes as a *repeated pattern* of behaviour, sometimes as a *kind of effect* in behaviour, and sometimes as the consequences of a specific mechanism that seems to make a difference in behaviour (for example, hormonal changes), depending on the explanatory interests. These do not map onto each other one-on-one, either. Consequently, as Longino argues, the different kinds of research, with different

explanatory interests, tools, and ways to understand the concept of behaviour itself will end up with incommensurable *explananda* – not only with different technical definitions, but also with actually different objects of research. This is not a problem in isolation, but it becomes problematic when we are interested in the *connection* between the subject matters (see Mitchel 1992), which is the case here.

Lacking generalizable criteria for a behavioural trait, we need to choose one that best fits the purposes at hand. I proposed evolutionary functionalism as an alternative way to approach behaviour in the previous chapter. This is not only a criterion for partitioning behaviour into traits for evolutionary purposes but also a perspective on how to integrate different explanatory dimensions and causal factors. It does so by connecting behavioural tendencies and environment (with an adaptive function), providing one sense in which to understand what the aboutness of behaviour is. There is a task that connects the parts of behaviour through their joint contribution to a positive fitness effect in which the parts are dependent on each other in making this contribution (as in the example of the honey buzzard's hunting behaviour). This does not require one-to-one mapping between a behavioural trait and a single mechanism or a module, or an identification of the trait with a particular set of bodily activities. This approach depends on a specific explanatory interest and is not a universally workable solution even in biology – and I am not suggesting that it is applicable to all human behaviour. It is, however, a sensible perspective from which to understand animal behaviour (for the reasons given in the previous chapters) and an adequate perspective on human behaviour when it is the evolutionary understanding of human behaviour that we are specifically interested in. There are limits to the applicability of evolutionary functionalism, as discussed before, and these are the limits of evolutionary human science. The object of my discussion is whatever remains within those limits.

The choice of perspective alone does not solve the problem of ambiguity. As we have seen before, the evolutionary perspective depends on assumptions about other explanatory dimensions. The

direct link between the function and behaviour made by the classical sociobiology does not work. The perspectives from different dimensions do not necessarily slice the behavioural patterns into traits in an isomorphic way, but different explanations may be relevant to each other or each other's presuppositions even if they ask different questions (see Mitchell 1992; Longino 2013). Moreover, whatever definitional perspective we use, the proximate causal basis of behaviour probably involves several cognitive and motivational mechanisms that are not restricted to this behaviour only (as we saw with the honey buzzard). Behavioural dispositions and capacities are psychological characteristics that are selected for the net effect they have on fitness through all the behaviour they participate in. This means that behavioural traits are connected to other behavioural traits through psychology, unless they are modular to a high enough degree to be completely distinct. Additionally, this means that *the psychological adaptations are adaptations only through their participation in evolutionarily functional behavioural traits*. The adaptivity of psychological traits depends on how they contribute to manifest behaviour. But the evolutionarily functional behavioural traits (the main *explananda* of evolutionary anthropology) and evolutionarily functional psychological traits (the main *explananda* of evolutionary psychology) cannot be mapped onto one-to-one relations. Rather, there is a co-evolution of adaptive behaviour and adaptive psychology.[112] This is a further reason to keep psychological and behavioural traits separate for evolutionary purposes: there may even be selective tension between them. Psychological development adds another connective factor between psychological and behavioural traits – but this complication is a topic for later.

[112] There are many reasons for why all psychology and/or behaviour cannot be completely adaptive, but the non-isomorphism of psychological and behavioural traits would probably be enough to make both psychology and behaviour sub-optimal. This is a further argument against adaptationism of most kinds, but not the holistic evolutionary functionalism proposed in the previous chapter since it is an instrumentalist idealization in the first place.

5.2. The Intolerable Ambiguity of Folk Psychology

There is no generally accepted definition for folk psychology. This is understandable since the definition depends partly on substantial views about what it is. For example, some consider it mostly a folk theory with applications while others think there is hardly anything theory-like in it. But as a starting point, I use the concept in an all-inclusive way, including cognitive capacities, interpretative practices, conceptual frameworks, quasi-theoretical ideas and other theory-like elements that exist in everyday conceptualizations, explanations and predictions of human behaviour, and in communicating and thinking about action and thinking. The folk theory part alone will be called *theory of mind* and the cognitive capacity to attribute mental states will be *mindreading*. None of these concepts has a fixed meaning in the debates and some of them are often used synonymously. Furthermore, it is not clear whether mindreading is supposed to attribute internal psychological states (it usually is supposed to) or whether attribution of mere agentive states qualifies as well, or what the attempted reference of the concepts in the theory of mind is, personal or sub-personal level states. Whatever folk psychology is, it is the starting point for both philosophical and scientific psychology, as well as theory of action and decision theory. In folk psychology, the agent, the subject of the action, is the causal locus of behaviour, and this makes it individualistic in its very presuppositions of the nature of agency.[113]

The debate over how folk psychology works as an interpretative device (is it a *theory* of mind that guides our perceptions and attributions or does it work in some other way?) includes both philosophical and empirical elements. There is another related debate over what the reference of its conceptual categorization is – the workings of our

[113] That is, this is the tendency in most of the discussion about folk psychology, both in philosophy and psychology. Whether folk psychology includes collectivist elements such as "we-intentions" as well (Tuomela & Miller 1988) is a further issue (see Tomasello 2009).

mind (which is at least partly experienced subjectively) or the modal states of the agent (epistemic and directive). Yet another question is the relationship between folk psychology and scientific conceptualizations of mind. These perspectives are central in the philosophy of mind and in the philosophy of psychology, but action theory and philosophical psychology are mostly interested in the logic of folk psychology *from within* the practice. There are two components in this philosophical enterprise: the explication and analysis of the essential conceptual elements of the theory of mind and the development of a more sophisticated theory based on them for further philosophical purposes, such as the analysis of moral responsibility. The latter includes normative issues in conceptualizing human action, such as rationality.[114] I will discuss some of these topics to argue for the distinction between agentive and psychological as a substantial distinction between two relevant ways to understand human action and its underlying psychology, not as two competing theories of intentionality. Moreover, I will endorse the *Pluralistic Folk Psychology* of Kristin Andrews (2012, 2015a & 2015b), the view that folk psychology is inherently pluralistic as a theory and as a practice, whether in the psychological processes involved in the practices, what its function in social life is, or what the references of its core concepts are. I will argue that the different levels of abstraction are not kept apart in everyday practices, and consequently some of the philosophical debate over the nature of human thinking and action may be confused in this respect too.[115]

---

[114] "Normativity" can refer to a wide range of issues involving "ought" rather than "is." The normativity of rationality is normative in a somewhat minimal sense: the rationality of an action tells us whether an action is *adequate* for the goals of the agent (see for example O'Brien 2019). I will return this later.

[115] Having elements of different kinds does not mean that there is no systematic connection between them. Georg Henrik von Wright (2001), for example, argued that "mind" refers to a complex interaction between (to paraphrase a bit) brain, folk-psychological conceptualizations of behaviour, and the conscious subjective mind. "Mind" cannot be reduced to any one component.

5.2.1. The Foundational Tension in the Philosophy of Folk Psychology

The strikingly different intuitions that have emerged in philosophy about the nature of folk psychology may be symptomatic of its pluralistic nature. Having strikingly different intuitions is not exactly unheard of in philosophy. Yet the disagreement of intuitions is not about which philosophical theory of folk psychology is correct but about what this philosophical theory should be about in the first place. We all have access to folk psychology from within. Starting from Wilfrid Sellars (1956), many philosophers have thought that folk psychology is essentially a theory – a folk theory that explains (or is used in the explanations of) human behaviour. This involves attributing mental states or entities, such as propositional attitudes, that explain behaviour and constitute the mind. Intuitively, this seems to hold: we not only theoretically explain but also predict others' behaviour based on mental attributions, and more importantly, we *manipulate* others'

Folk psychology fixes the semantics of the mental categories, brain is the source of causal powers, and reflection provides a privileged epistemic access to the complex that allows us to know intuitively what we are doing (i.e. what is the action we intend to take). Even without going this far, one could argue that folk psychological practices nevertheless play a constitutive role in what the human mind is – either conceptually (as in von Wright's image) or causally (through *mindshaping* (Mameli 2001)), or both. I do not have a stance on what the reference of "mind" is, nor do I think it matters for any of the issues at hand as long as the local references are clear. Instead, I try to avoid using the terms "mind" and "mental" except as assumptions within the theory of mind as an object of discussion. I will use "psychology" and "psychological" as a narrow category of sub-personal processing. These processes may depend on interaction with the environment, and I am sympathetic to attempts to expand the perspective outside strictly internal processes, but I will use the "psychology" of internal cognitive processes as a contrast to both the subjective conscious mind only and folk-psychological ascriptions of agency with mental states. I use "psychological" rather than "cognitive" because psychological drives relevant to explaining behaviour are not cognitive but motivational states.

future behaviour by manipulating their mental states through persuasion, arguments, threats, *et cetera*, for example. In other words, mental states, whatever they are, seem to work as if they are causal factors, according to the Woodwardian understanding of causality. This means that they cannot be mere *descriptive abstractions*, but they have to refer to something that causes the behaviour or, at least, to causal processes that ground these abstractions. At the same time, however, the propositional attitudes are *intentional* (directed to something, being *about* something; Anscombe 1967; Dennett 1971) and *rational* (that is, they are connected through the entailment relations based on their propositional content). It seems that this is foundational for the semantics of folk psychology, whilst the underlying causal processes produce the behaviour itself (Davidson 1963 & 1970; von Wright 1971 & 2001; Kim 1993 & 2005). Much of the philosophy of mind for the past half a century or so has been about various solutions to this apparent dual nature of mentality. I will not go into any of these debates more deeply than necessary.[116] But I do have to discuss some issues to

[116] I am only interested in individuating mental states as explanatory factors. To do so, I have to say something about mental causation, rationality in explanations, folk psychology, and mindreading, even if I am not going to go deep into any of these issues or try to make arguments within these debates. But I will not discuss *mental content*. There is a whole subfield of philosophy of mind devoted to whether the right way to understand the contents of beliefs and other mental states is to restrict oneself to the internal properties of individual such as representation (*narrow content*; for example, Fodor 1987 & 1990) or whether the content extends to facts about the world (*broad content*; for example, Putnam 1975 & Burge 1979 & 2003). This is related to the issue at hand through an overall view about what mentality is about. The idea of narrow content is connected to the idea that mental attributions are about attributing psychological states that have a narrow content. The view that mental attributions are ascriptions of relational states between an individual and external facts is related to broad content theories. Some of the more recent views about mentality, such as *enactivism* (for example, Hutto & Myin 2013 & 2017), attempt a radical reinterpretation of mentality on both accounts. There is no necessary conceptual connection between the two, however. There is a

articulate properly the adequate *explananda* of the evolutionary explanations and what the assumptions behind this are.

The *representational theory of mind* is the dominant paradigm and I assume that something like it is true when it comes to the cognitive-level scientific descriptions. Whether the representations postulated by this theory are mere *instrumental abstractions* of the functional properties of the cognitive architecture in modelling the performative capacities of the human brain (and are, therefore, emergent systemic properties) or *entities* with inherent causal powers that are implemented by the brain processes (and are therefore *parts* of the system) is irrelevant here. The core of the theory is that the mental entities are intentional yet causal at the same time. This usually involves reducing the agent-level intentionality to psychological-level representations that participate in the causally effective cognitive processes (for example, Fodor 1975, 1981 & 1990; Shea 2018).[117]

---

difference between how to understand a mental state as a container of a mental content (the person as a whole or a specific psychological state) and whether to understand the content in terms of the individual's states or include external factors. For example, whether the meaning of "water" includes facts about the chemical structure of water does not depend on whether we think beliefs about water are representational states in our cognition or dispositional states of an individual. It depends even less on how folk psychology works as a conceptual framework for presenting explanations. There are overall philosophies of mind that draw borders regarding several issues and the themes are connected, but not through direct entailment. Furthermore, if my distinction between psychological and agentive is accepted, the only direct consequence for the debate over mental content is that it makes a hybrid position available: cognitive (representational) contents and individual (doxastic) contents are kept conceptually separate and their substantive relationship is left open.

[117] The representations have content, but their relation to belief-like states and to propositional contents is an open question. The core property of representations is that they stand for something in cognition without being in direct causal connection to the object; they are decoupled from the immediate presence (see

The *eliminativists* think folk psychology is a *false* theory, and the mental entities that it attributes (and the mind) do not even exist (Stich 1983; Churchland & Churchland 1998). Patricia S. and Paul M. Churchland (P.M. Churchland 1981 & 1989; P.S. Churchland 1986; Churchland & Churchland 1998) conclude from this that the adequate level of scientific explanation is neuroscience, and that scientific psychology is based on the same false theory as well. Stephen Stich argued in his 1983 book, in contrast, that the level of descriptions of cognitive science is fine – but without the intentionalist baggage from folk psychology (see also Cummins 1983 & 1996).[118] But not everyone agrees that folk psychology even attempts to describe the psychological reality of cognition. The *non-causalists* (for example, Anscombe 1967; von Wright 1971) concluded early on that folk psychology is not a theory that postulates entities at all, and it does not give causal explanations for behaviour – it is about rationalizing the action of an individual agent, a person, by attributing reasons for action that explicate what the individual's behaviour is about. Some causalists (Davidson 1963 & 1970; Dennett 1971 & 1987; Lewis 1972) had similar insights, proposing that the rationalizations of action are ascriptions to human agents as "systems" that have real causal properties, but the causal properties are captured under different descriptions than rationalization.

This early disagreement on the very function of folk psychology (that is, whether it refers to causally effective internal entities or ascribes mental states to human agents as holistic systems) may be symptomatic of the vagueness of reference (see also Stich 1996). These approaches do not need to be rival views of the *same* thing. They may be attempts to capture different aspects of human agency and psychological phenomena related to it that are unified in folk psychology –

Sterelny 2003). Hence, they function as having mental content even if they are a part of a causal system (see Shea 2018). I will return to some of this later.

[118] Later, of course, Stich took an agnostic position about the reference of folk-psychological terms (Stich 1996).

and they may be unified at the level of detail that is necessary and sufficient for folk-psychological practices, but not for further sophistication. This is in large part an empirical claim, in two ways: how the mind works and how interpreting it works (or, what interpreting minds is about). These are separate but closely connected questions. If our use of folk psychology includes assumptions about mind, these assumptions are a part of our interpretation of the object of folk psychological interpretations when we reflect on our intuitions about the reference of folk psychology. If we want to discover the possible biases in our intuitions caused by folk psychology, including about folk psychology itself, we need to understand the psychology of *mindreading* (the attribution of mental states; see Baron-Cohen *et al* 2000; Nichols & Stich 2003; Apperly 2010): how, based on what and to which purposes we make these attributions. The philosophical analysis should be amended by empirical knowledge about this.

5.2.2. Psychology of Folk Psychology

Psychologists and primatologists generally use the term "theory of mind" when talking about the capacity to attribute mental states to others (for example, Gopnik & Meltzoff 1997; Wellman & Liu 2004; Call & Tomasello 2008; Lurtz 2009; Henry *et al* 2013), but it is not clear how strong their assumptions of theory-likeness are. Generally, they seem to merely refer to the ability to attribute mental states to other beings (that is, mindreading). Some psychologists have argued for much stronger theory-likeness, based on the observation that children seem to use natural language as a theory-like guide to human behaviour (for example Gopnik & Meltzoff 1997; Wellman & Liu 2004). The psychological vocabulary seems to refer what is "behind" behaviour.[119] Others think that folk psychology, or its core at least, is innate

[119] Bogdan (1997) and Hutto (2008), on the other hand, discuss the role of language at later stages of acquiring full-scale folk-psychological capacities through learning narratives. Narratives structure children's thinking about

rather than acquired with language, but it functions as a theory of other minds (for example, Baron-Cohen 1995; Carruthers 2013b). Some psychologists also make a distinction between *affective* theory of mind (the ability to understand feelings and emotions) and *cognitive* theory of mind (the ability to attribute beliefs, goals, and intentions) (Shamay-Tsoory *et al* 2010; Duval *et al* 2011). This is an important distinction as the two capacities might function very differently. An alternative to the *theory theory* of folk psychology is the *simulation theory* (Gordon 1986; Goldman 1989 & 2006). The directions of inference are opposite in these theories: in simulation theory, mindreading involves generalization from internal observations. The simulationists think that interpretation has more to do with mental simulation of other people's context and trying to get "into their heads" with introspective knowledge about mental states that would emerge in the given situation.[120] This does not involve theoretical postulates but cognition involving introspection. Lately, hybrid theories have become more popular (see Nichols & Stich 2003; Perner & Kühberger 2005). All these views are intellectualist views of the mindreading process, and there are those who argue that much of folk psychology is something else (for example, Bogdan 1997; Hutto 2007 & 2008; Hutto et al 2011; Zawidzki 2013; Andrews 2012; 2015a & 2015b).

The psychology of mindreading is concerned with how the attribution of mental states works as a part of human cognition. This is connected to two philosophical issues: the epistemic justification of making claims about other minds, and whether folk psychology involves representing mental entities (that may or may not exist) that are psychologically explanatory for action. With respect to the epistemic issue, the simulation theory is an account of the logic of mindreading that bypasses the problem of how to *test* the hypothesis of

---

action as an event that is a part of a chain of events involving beliefs acquired earlier and goals in the future.

[120] The simulation theorists do not have a unified view of what counts as simulation (see Gallagher 2007), but this is not relevant to the current discussion.

other minds (since this attribution is not hypothetical). However, it is not sufficient to account for how we perceive our *own* mental states without some sort of representation of mind (Carruthers 2009 & 2011). This means that even if the simulation theory is correct about how we attribute mental states to others (in contrast to the theoretical hypothesizing of theory theory), it does not rid us of a theory of mind as a conceptual framework that structures the internal observations (see also Dennett 1991a). The philosophical issue about their reference (psychological or agentive, and real or fictional) remains, even if the epistemological problem of other minds is fundamentally transformed. Furthermore, simulation would not make the attributions by mindreading necessarily correct (see Stich 1996), and how we perceive our own minds (which is the basis of knowing other minds in simulation) may also be fundamentally misleading in the first place.

Observing our own minds involves the same folk-psychological framework we use in interpreting others in mindreading, regardless of which one is the psychologically primary context for it. But introspection involves internal access to some of the psychological processes involving how we represent the world, what our needs (or drives) and objects of inclinations are, and what our current action is about, producing conscious *metarepresentations*. Both philosophers (James 1884; Ryle 1949; Boghossian 1989; Carruthers 2009 & 2011; Doris 2015) and psychologists (Nisbett & Wilson 1977; Nisbett & Ross 1980; Hurlburt & Schwitzgebel 2007) have seriously challenged the epistemic certainty of knowing one's own mental states. At the same time, it would seem absurd to deny that we know what we are doing most of the time. For example, it may be a matter of interpretation for others to know whether my hand-waving is about greeting someone, stopping a cab, or getting rid of a wasp, but I can hardly be mistaken about this myself under normal conditions. But we need to make a distinction between knowing one's own intentions and knowing the psychological processes underlying it. We may not have access to all the decision processes, for example. Nevertheless, we may grant enough reliability and directedness to introspective observations to

give it a privileged (although fallible) epistemic status (see Nichols & Stich 2003; Goldman 2006; and Carruthers 2011 for such accounts) and yet recognize that the conceptualization of these observations is theory-laden. We cannot simply assume that the folk-psychological interpretations of our own minds are raw observations, but observations with a conceptual structure. Regardless of the primary reference of these concepts (inner psychological states, agentive dispositions, or a heterogenous category), the framework (also) serves the purpose of positioning the agents (including the subject of the interpretation herself) in a network of relations to the world. Beliefs and desires, whether they have psychological references or not, have a propositional content that describes these relations, epistemic or directive. Even if beliefs and desires have psychological realizers (representations), the semantics of folk psychology depends on these holistic relations.[121] On the other hand, since we have access to our own minds, it would be curious if we did not also take advantage of this in inferring about other minds. This may be an effective cognitive tool, but also a basis for confusion if it makes the reference of the concepts a heterogenous category.

Concentrating on the most sophisticated levels of the mindreading process can, however, be misleading. Our mindreading cognition probably uses more direct and minimal ways of attributing perceptual states and behavioural dispositions to others than a full theory of mind in much of our social practices. The capacities for this task include joint attention, mimicry, and affective empathy, for example, regardless of which theory turns out to be the correct theory of more sophisticated attribution. (See Sterelny 1995, 1998, 1999 & 2003; Bogdan 1997; Tomasello *et al* 2003; Hutto 2007; Gallagher 2008 & 2012; Tomasello 2008; Zahavi 2008; de Waal 2009; Hutto *et al* 2011; Andrews 2012, 2015a & 2015b; Zawidzki 2013; and Hutto & Myin 2017 for a

[121] There are also philosophers who have suggested that the semantics of folk psychology is even more intimately connected to behaviour (Ryle 1948; Wittgenstein 1953; von Wright 1971 & 2001).

discussion of the lower-level cognition involved). Kristin Andrews (2012, 2015a & 2015b), for example, has proposed a theory that she calls *Pluralistic Folk Psychology*, highlighting both the variety of psychological mechanisms involved in folk psychology (mindreading being just one) and the functions performed by folk psychology. It is not just about attributing propositional attitudes. Action explanations and predictions may rely on other factors, such as personality types, social roles, context, and so on, and the functions of folk psychology include interactive practices like coordination and justification. Moreover, explanation (the focus of much of the theoretical discussion) and prediction (the focus of most of the empirical research) may involve different capacities. She argues that prediction relies mostly on the *normative* aspects of folk psychology that regulate social interaction, and only the *violation* of these norms evokes the need to explain what happened. This explanation may involve attributing cognitive states, emotions, or personality traits, for example, which loops back into further predictions.[122] Some of the attributed traits may be on the personal and some on the sub-personal level.

Furthermore, even if the pure theory theory were the correct way to think about the logic of sophisticated mindreading (as postulating explanatory entities, events or processes for external behaviour and its directedness, without internal observations of these entities), the full theory of mind would still be *preceded* by a more minimal theory of mind, both developmentally and evolutionarily speaking. We should probably talk about interpretive capacities with quasi-theoretical presuppositions rather than a theory of mind in these early stages, even if we espouse theory of mind. Given how psychological architectures are usually constructed, more sophisticated mindreading tools

[122] This is well in line with empirical studies that suggest that people rely on behavioural expectations rather than theory of mind in natural contexts and that the inferences using theory of mind are cognitively taxing, deliberate, and, as a result, adult humans make often mistakes that they are equipped to avoid. See Malle & Pearce 2001; Keysara et al 2003; Apperly et al 2006, 2008 & 2010; Bryant et al 2013.

are more likely to be additions to than replacements of the earlier stages. The central empirical question of philosophical consequences is what the *function* of these capacities may be. To address this question, we may need to expand the discussion to evolutionary origins. The selected use of the capacities during the main stages of their evolution tells us the success conditions of their use – that is, what is the object that they should most accurately trace in interpretation. Given the usual constraints of adaptationism, we should not expect them to be completely successful in this or that the answer should be clear, but we can get some idea of the reference of the concepts in our practice.

5.2.3. The Evolution of Folk Psychology

The evolutionary background and function of this set of diverse capacities and practices is important for the following reason. If we approach the core concepts of folk psychology as quasi-theoretical terms, we must ask what entities they are "trying" to refer to, what their function in the evolved practice is. In other words, the practices are attempting to trace something in other individuals, and the central concepts used in the practice get their reference as a part of this process.[123] If the evolutionary function of folk psychology were to correctly refer to the psychological structure of mind, its accuracy in this would be selected. But if folk psychology is a set of capacities and practices in social interaction and its evolved function is more pragmatic, its quasi-theoretical components do not necessarily serve the

[123] This may be understood in the terms of teleosemantics in that the function of the representations in the cognitive processes involved is to trace certain features, and they refer to what those features are (Millikan 1982 & 2004; Papineau 1984; Dretske 1988). However, the point here is independent of any theory of reference. Any causal-historical theory of reference needs a story of the "baptism", which will need a description of what the foundational practices were, and any descriptive theory needs to describe the function of the concepts in the folk theory of mind. They all will lead into the same question.

purpose of *describing the structure and functioning of mind* (as something internal). Instead, they conceptualize the *states of the agents in relation to the world* in a way that makes prediction of their behaviour possible and facilitates social interaction.[124]

The evolutionary origin of folk psychology and its various elements is commonly thought to lie in enabling and boosting our wide variety of social practices. We use it to predict others' behaviour and signal our own intentions in contexts of collaboration, competition, manipulation, moral evaluation, and so on. (See Vygotsky 1978; de Waal 1982; Byrne & Whiten 1988; Byrne 1997; Bogdan 1997; Corbalis & Lea 1999; Richerson & Boyd 2005; Tomasello *et al* 2005; Knobe 2007; McGeer 2007; Moll & Tomasello 2007; Tomasello 2009; Emery 2012; Devaine *et al* 2014; Andrews 2015b.) To be successful in this, the details of the cognition do not necessarily matter. For most tasks related to sociality, robust representations of the relations between the interpreted individual and their environment are sufficient. But it is also likely that growing sociality and new cognitive needs create selection for *different* capacities, not just more powerful or more accurate ones.

Radu J. Bogdan (1997) has built a whole theory of different cognitive stages in mindreading (both evolutionary and developmental), and what cognitive tasks the various stages accomplish, evolutionarily speaking. The stages range from "natural teleology" (understanding the directedness of behaviour and different perspectives) through a "psychobehavioural" stage (attributing non-visible goals and context-independent epistemic states to the agent) to a "psychosocial" stage (with higher abilities that are needed in two-directional social interaction). Bogdan's theory is highly speculative at times, especially when it comes to the evolutionary dimension, but the details of the

[124] "Evolutionary debunking arguments" are sometimes presented for scepticism on various issues related to the products of human cognition on the basis that nature selects for usefulness, not truth. But the usefulness is often dependent on correctness of these products, depending on the case. (See Wilkins & Griffiths 2012.) The point here is not about debunking, but about the locus of what needs to be got right.

evolution (or development) of mindreading are not important here. The important idea is that the rudimentary mindreading practices do not *need* a theory of mind in the sense that it postulates internal psychological states. The social cognitive capacities of small children and apes could be achieved without such. (See also Hutto *et a*l 2001; Hutto & Ratcliffe 2007; Andrews 2012).[125] The more recent elements of social cognition, in turn, may involve both attributions of psychological states and locating the targets of interpretation in narrative structures involving the rationalization of past events and future goals (see Bogdan 1997; Hutto 2008), which are expansions of two different kinds, cognitive and agentive (cf. Goldie 2007).

Whether or not non-human apes have a true theory of mind is contested within primatology. Chimpanzees seem to understand the difference in others' perspectives and goals, to some degree, and use this information in social contexts, and this is usually interpreted as their postulating an inner life to each other (see Call & Tomasello 2008), but primatologists Daniel Povinelli and Jennifer Vonk (2003) have questioned this (see also Penn & Povinelli 2007; Andrews 2012). Instead, they suggest, chimpanzees might merely use *behavioural abstractions*. This is in line with Bogdan's interpretations. In addition, they accuse the more generous interpretations of *using* the theory of mind in interpreting the "interpreting" in chimpanzees, instead of *discovering* it in them. That is, the researchers are using their own mindreading capacities in reading mind on the chimpanzee being observed by another chimpanzee, and then again when they interpret what the observing chimpanzee "must" be observing. This is a nice example of the folk-theory-ladenness of observation that might bias

[125] Whether small children are already using a more sophisticated theory of mind is a different question to whether they need it to accomplish the tasks they perform. Peter Carruthers (2013b) argues that they do use more sophisticated cognitive tools early on. But this may be telling of the need for a sophisticated theory of mind in adult age, which necessitates the development of one. The developmental and evolutionary stages are often correlated, but do not completely mirror each other.

observations of human behaviour as well. But the important thing here is not whether chimpanzees really attribute psychological states (see Tomasello & Call 2006; Sober 2009; Butterfield & Apperly 2013; and Clatterbuck 2015 for a defence of a stronger interpretation). The important idea is, once again, the distinction between attributing behavioural abstractions and attributing psychological states, and the difficulty of knowing which of these is the case, based on chimpanzees' performative competence and behaviour alone. In the first case, the chimpanzee under interpretation is the locus of the dispositions, as an agent, and this would be enough for chimpanzee social cognition. A folk psychology as an expansion of this would be a more sophisticated conceptual framework to locate the individual in the environment with behavioural dispositions, transcending temporal and spatial limitations.

If this were the case of ape cognition, it would probably also be partly true of human folk psychology. We clearly make assumptions about internal states, too, but there is no reason to think that human mindreading that uses psychological attribution has *replaced* all the earlier elements in cognition after emerging. This is not how evolution works – quite the contrary.[126] Furthermore, children have some folk-

[126] It is not likely that the evolution of mindreading capacities begins with attribution of inner states, or that our more sophisticated cognitive capacities have completely replaced the more rudimentary ones. A nice example of layered cognition like this is Frans de Waal's (2007) famous "Russian doll model" of empathy, a model of how perception, cognition and motivation regarding others works in a system where cognitively more sophisticated capacities are new "layers" built on the older ones. The core mechanism is a direct emotional connection between the other's behavioural state and the subject's states (automatic motor mimicry and emotional contagion). This provides the basis for a higher layer in which the other is perceived to be the source of felt emotions (and the object of care), which in turn is a basis for the fully developed empathic layer of attribution and perspective taking. Each earlier layer plays a role in the higher layers. The general lesson from the evolution of mind is that we should expect it to be layered like this in its capacities, including in the capacity to interpret other minds.

psychological understanding before they have any concept of belief (Wellman, Cross & Watson 2001), and mindreading and applying it in inferences develop in stages of cognitive competence (Wellman 1990; Apperly & Robinson 2003; Moshman 2004; Henry *et al* 2013). In fact, there is evidence that even adults attribute beliefs and make belief-related inferences deliberately, not automatically, and much of folk-psychological practices work without it (Malle & Pearce 2001; Keysara *et al* 2003; Apperly *et al* 2006, 2008 & 2010; Back & Apperly 2010; Bryant et al 2013).

However, Stephen Butterfield and Ian Apperly (2013) have introduced another alternative to interpreting rudimentary interpretation capacities as having a *minimal theory of mind*.[127] Under this interpretation, chimpanzees and young children do not attribute mere behavioural dispositions but refer to the underlying psychological states (something that is in between the observed environment and the behaviour) in their interpretations in some robust way (see also Whiten 1996). But they do not use a full theory of mind, either, since they lack metarepresentational categories in their cognition. They do not have beliefs about others' beliefs, but their cognitive processes represent the cognitive processes of others. This is definitely a theoretical possibility when it comes to accounting for things like automatic attributions of visual perceptions and immediate goals of behaviour to others. However, the minimal theory of mind is precisely not about attributing propositional attitudes. If the richer theory of mind is an expansion of a minimal theory of mind like this, it requires additional attributing of world-directed states that are the *person's* attitudes, with propositional contents that are linked to the person's other propositional attitudes through the contents. The success criterion for this is still whether the attributions relate the individual to her or his environment in ways that serve social purposes – that is, prediction, coordination, *et cetera*.

[127] Butterfield and Apperly do not argue that this is what actually happens in chimpanzee interpretation but introduce the notion of minimal theory of mind as an alternative way to think about this and various other cases.

If the evolution of cognition is tied to the evolution of social practices, there will be selection for isomorphic relation between the cognitive processes and their interpretation by others, either directly, by *mindshaping* (Mameli 2001; Zawidzki 2013; I will return this shortly), or both. But this selection is not directed to make the enriched interpretation of the internal processes more correct, but to have a robust connection between the two. Let us bring back Beatrice saving Amos from the earlier example. A minimal theory of mind would attribute her a state of perceiving Amos in distress and a teleological state of mind to help Amos. But how does this change if we attribute her a belief that Amos is drowning or the pro-attitude to save him? These still seem to be attributions of the overall states of Beatrice to her. If the theory of mind is based on a minimal theory of mind, it is still an open question whether attributions of propositional attitudes can be understood in terms of psychological-level representations. They may. However, if Pluralistic Folk Psychology is true (and it seems to be compatible with the empirical psychology of folk psychology and evolutionary considerations) it would make sense to think that folk psychological practices generally rely on more robust attributions of intentional states to a person under interpretation but go "deeper" into the thinking processes of the individual when it is needed. For the practical purposes of folk psychology, it does not make sense to distinguish between overall assessment of what is intended and why, and going into more detail, even though this might involve a transfer between levels – and constitute a category mistake, strictly speaking. Even if the function of mindreading is the agentive level dispositions, it can still involve attribution of internal states as a part of this process. Since we humans have access to our own minds (even if sometimes fallible), it only makes sense that we use our self-observations as a clue to what is happening with other agents, too.

All this is somewhat speculative, and it could not be otherwise – the debate on the empirical issues has not been concluded yet. However, my argument is only that the two levels are both conceptually and substantially separate, that it makes sense to talk about both, and

folk psychology is about both but without maintaining the distinction. Both options for explaining the primitive interpretation (behavioural abstraction and minimal theory of mind) implicitly make the distinction between psychological and agentive, although they propose different levels as the object of rudimentary interpretation.

If there is no need to distinguish between different levels (agentive and psychological) in folk-psychological practices, however, the levels conflate with no practical consequences. If there is no selection for accuracy, mixing the levels may even be cognitively less demanding and thereby more functional. We also think about our own future actions and plan for them, which involves reflection and reasoning, which in turn involves conscious rational thinking and direct knowledge of what we are planning to do. But even here the function of folk psychology does not lie in accuracy in representing our inner processes. The function of folk psychology is not so much a theory of mind as a theory of agency that makes presuppositions about mind and utilizes introspection in this. There is no reason to think that folk psychology's *principal function* is to describe the inner workings of mind. At the same time this alone does not entail that the attributions (such as beliefs and desires) are not *also* attempts to capture internal psychological states. The emergence of the theory of mind does not need to create another layer in our mindreading practices. Using the theory of mind in interpretation may instead be "thinking harder" about the beliefs and goals of the agent, rather than attributing further explanatory entities behind the agentive states. This would be more parsimonious on the level of robustness needed for folk-psychological practices. As for philosophical and scientific endeavours, whatever entities, processes, and structures constitute human psychology, we can think about them as the realizers of agent-level beliefs and desires, or their constituent parts. But this is a further question. Folk psychology's success criterion is pragmatic: it must be useful, and for this it is

enough that it describes the robust dispositions of an individual agent, not accurate details of its underlying psychology.[128]

To sum up our discussion of the empirical side of the issue, folk psychology seems to involve a variety of quasi-theoretical and cognitive tools that we use to understand others' behaviour in a certain context (see also Davis & Stone 1995; Bogdan 1997; Baron-Cohen *et al* 2000; Nichols & Stich 2003; Hutto & Ratcliffe 2007; Hutto 2008 & 2009; Andrews 2012, 2015a & 2015b). It involves things such as perspective taking in a context; attributing belief-like states that describe the agent's cognitive relation with the world; attributing immediate and non-immediate goals to behaviour; simulation and empathetic production of similar mental states; conceptualizing the behaviour of the observed agent as action with agentive states such as beliefs, desires, goals, intentions, *et cetera*; attributing broader "mental backgrounds" such as belief systems, long-term goals, emotions, moods, personality traits, *et cetera* to the agent; and ideas about the behavioural tendencies that humans generally have. The more sophisticated stages of folk psychology may involve symbolic learning, such as language and explanatory narratives (situating the target of interpretation into a series

[128] This view about folk psychology also has deep roots in philosophy. It has been popular lately among those who approach the issue from an evolutionary point of view, and these are the theorists referred to here. But the idea that mind-related concepts obtain their semantics from understanding action, especially other people's actions, instead of referring to mental entities in some more concrete way, goes back to Ludwig Wittgenstein (1952) and philosophers under his influence (see Ryle 1949; Anscombe 1967; von Wright 1971 & 2001). There is also a close connection between interpretation of action and mental states for the likes of Donald Davidson (1963 & 1970). Wilfrid Sellars (1956), who originated the idea of theory of mind, also considered the actions of others to be the starting point for attributing mental states (that he was realist about as internal states, though). Later, philosophy about these issues branched into philosophy of action and philosophy of mind, one of which is about understanding action and the other about human psychology. They both keep using the same folk-psychological framework, but the objects of analysis may have diverged.

of events, actions, and goals; see Bogdan 1997; Gopnik & Meltzoff 1997; Hutto 2008). Its basic elements, as well as the tendency to learn its acquired elements from others, are deeply entrenched in the development of human psychology and have evolutionary roots. Other animals share some of these capacities, but not all. (See Bogdan 1997; Tomasello *et al* 2003; de Waal 2009; Andrews 2012; Tomasello 2014.) What matters for the main topic under discussion here is how folk psychology functions as an explanatory practice and what its relationship to proper psychological explanations is.

5.2.4. In Search for Clarity (by Making Things More Complex)

The most plausible interpretation of both the empirical research and the existence of differing reasonable intuitions within philosophy is that the reference of folk psychology is pluralistic and simplifying. Kirstin Andrews (2012, 2015a & 2015b) has argued for a similar position, as discussed above. Pluralism also implies pluralism in the explanatory status of folk-psychological attributions. Peter Goldie (2007) has also argued that folk psychology mixes different elements: rationalizations and narrative-historical explanations, both of which belong to agentive explanations in my distinction, but also explanatory references to non-rational (even irrational) motives, emotions, moods, personal characteristics, *et cetera* that I will consider psychological causal explanations. He argues that philosophers should not think of folk psychology in terms of rationalizations alone, and that the project of building philosophical psychology based on this is fallacious. I agree with the first point about how to think about folk psychology, but I think that agentive descriptions are a specific way to understand action (within and without folk psychology) and there is a role for philosophical psychology based on rationalizations for as long as it is understood to be a conceptually different project from scientific

psychology.[129] However, the issues about the psychological makeup of humans and the theory of mind (as a conceptual part of social practices) should be analytically divorced. Among other things, this would make the eliminativist arguments philosophically irrelevant to the folk-psychological theory of mind.[130] Moreover, it raises a question about the role of the theory of mind in *explaining* action, among some familiar lines that I will discuss in the next section. Furthermore, these levels connect more intimately than "mental" and "physical" levels, as it were.

Moving from folk psychology to "mind" itself, it would be tempting to follow those who distinguish between three conceptually different ways (or "levels" of abstraction) to talk about human behaviour and its causes.[131] First, the *neurophysiological* level, referring to physical

[129] I disagree with Donald Davidson's (1974) stance that empirical psychology, decision theory and philosophical action theory lie on the same continuum. I generally agree with his point about scientific and philosophical issues lying on a continuum and being relevant to each other in a symmetric relation, but in this particular case, this is the wrong way to relate them.

[130] Eliminativism, under this view, is mistaken about the function of folk psychology, the reference of its concepts, and how successful it is. This was pointed out early in the discussion on eliminativism (e.g. Horgan & Woodward 1985; Bennett 1991; Dennett 1991b; Bogdan 1991b & 1993). Patricia Churchland (personal communication), has agreed that it is possible (even probable) that our social practices necessarily require a theory of mind and its categorizations, given how our cognition works. Her issue is explicitly with the scientific understanding and its postulates. Steven Stich, on the other hand, soon turned to a sort of agnosticism about the whole issue (Stich 1996) and has more lately focused on the issue of folk psychology as a psychological capacity (i.e. Nichols & Stich 2003).

[131] These ways are applications of three different conceptual frameworks that aim at different ends. There is no need for any metaphysical implications in making this distinction.

processes only.[132] Second, the *psychological* level proper, which gives a functional, computational, or algorithmic description of these processes (Putnam 1975; Fodor 1975, 1981 & 1990; Marr 1982; Cummins 1983; Shea 2018). This would be the level of the representational states that cognitive science studies and that give rise to the psychological phenomena (such as abilities and inclinations) that the psychological science describes. The descriptions on this level refer to causal processes and dynamics between representations, drives, emotions *et cetera*. This is also the proper level of evolutionary psychological explanations. Third, the *agentive* level, which is the level of attributions such as beliefs, desires and intentions, understood as states of the whole agent. In other words, "mental states" would be divided into psychological (in a narrower sense) and agentive, and the debate over the nature of intentional states would turn out to be partly a confusion between different levels instead of a substantial debate only. I will follow this temptation after discussing some complications.

Distinguishing between cognitive architecture and agent-level descriptions while regarding them both as valid perspectives is an old idea when it comes to the actual analysis of how the mind works. The view outlined here is similar to Daniel Dennett's idea of intentional, design, and physical stances (Dennett 1987), for example, and one standard solution to understanding intentional states is precisely that they are useful ascriptions to the whole person, as mentioned before: they are states of the agent, not references to separate internal states (Davidson 1974; Dennett 1987). Distinguishing the levels, ascriptionism becomes compatible with a causal interpretation of psychology proper. The ascribed states are just properties of the system, rather than parts of the system, and thinking of them as parts with causal role would be a category mistake. To quote Dennett's (1991b)

[132] A proper "physical" level would be yet another level, however. Neurophysiological processes are functionally organized and not theoretically reducible to physical processes only, but this is irrelevant here.

metaphor, beliefs and desires are more like the centres of gravity than the concrete states of mechanisms.

There have also been attempts to build a theory of cognitive representations explicitly without any connection to the teleological notions of folk psychology (for example, Stich 1983; Cummins 1983 & 1996). These theories are usually considered to be alternatives to the representational theory of mind (Fodor 1975, 1981 & 1990; Shea 2018) that attempt to explain intentionality in terms of cognitive-level entities and propose that intentional states of the agents are representations within the system. What I suggest, however, is that both agentive and psychological level are sensible levels of analysis, even if we also understand intentionality and representations at the psychological level. The latter is a separate question asking *what explains*, on the cognitive/psychological level, agentive-level intentionality – if anything does. Furthermore, neither is the (only) valid level of folk-psychological references. Philosophical theories of mind are not explications of folk psychology but revisionist theories.

Some of the recent naturalistic attempts to make sense of intentional representational mind approach the processes of mind as brain processes with robust outcome functions that have been stabilized by evolution and learning (for example, Godfrey-Smith 2006; Sterelny 2015; Shea 2018). These processes are controlled by sub-personal subsystems, and their functional operations have representational content. Personal-level attributions of beliefs and desires, however, are robust states of the individual (or the whole "system") that describe their cognitive relations with the world, descriptive and directive, and these relations are constituted by the parts of the representational system. This is plausible and I will not challenge it as such. However, it would still be a category mistake to reduce the states of the system to the parts of the system. Beliefs and desires are dependent on the system of representations, not parts of it. If the robust outcome function approach is correct, it explains the constitution of systemic states, but this is a different matter and does not build a conceptual link between the levels. There are also interesting alternatives to the representational

theory of mind within the naturalistic context that are still compatible with describing humans as agents. The most extreme is *radical enactivism* (Hutto & Myin 2013 & 2017), which takes the biology outside the central processing system more seriously as part of cognition. It proposes that much of cognition lacks any representational content at all and has more to do with how the sensory-motor system functions as a whole. The so-called "4e movement" (enactive, extended, embodied and embedded; see Newen, De Bruin & Gallagher 2018; see also Clark & Chalmers 1998) approaches to mind and cognition in general challenge the classical representational theory of mind, without necessarily denying that representation is a part of the picture.[133] But even if they are right, it would not directly make intentional action descriptions inadequate on the agent level. The debate between representationalists and their critics is not about attributing mental states to agents but about how the mind works. The latter includes the issue of what *explains* the applicability of folk-psychological attributions to human agents. There are several ways to do this, including representationalist intentional realism (reducing agent-level intentionality to cognitive processes) and instrumentalism (intentional stance, and ascriptionism in general). This part of the question is mostly irrelevant to the issue at hand and will be left aside.

Furthermore, a precise relation between psychological and agentive levels is more difficult to understand with directive mental states than with descriptive ones. Psychological-level descriptive representations and agentive level beliefs can be thought of as being in a complex constitutive relation, for example.[134] But how the drives and

---

[133] Some proponents of the representational theory of mind (most notably Fodor 1975, 1981 & 1990) are quite explicit in reducing individual-level propositional attitudes to mental representations (being *intentional realists*), but this is a further claim that a proponent of the theory does not need to make.

[134] Even here there are complications. It seems that people sometimes act as if they believe something, even if those beliefs are not a part of their deliberated or conscious belief-system and might even be doubted by the agent if they become aware of them. There are various attempts to differentiate between

motivational salience relate to agentive-level desires and other pro-attitudes is trickier. Motivational salience is a crucial explanatory component in the emergence of pro-attitudes, but it is difficult to see how it alone could have the right kind of propositional content. It is simply a causal factor, incentivising or aversive, that instigates behaviour. Motivational salience explains preferences but is not itself a preference with content. Furthermore, pro-attitudes are about particular goals, not behavioural tendencies towards or away from a behaviour, which motivational salience entails. The goals implied by a pro-attitude may be quite general and abstract, of course (like "world peace", being famous, or whatever goals moral values entail even prior to knowing these entailments), and folk psychology accommodates *moods* and *personalities* as more general and robust dispositional states. But these are not the same as a tendency to be motivated by certain things in certain contexts. This is evident in how folk psychology is inadequate in capturing mental episodes such as depression.

---

stronger and weaker notions of belief, such as distinguishing between beliefs as involuntary dispositions to feel that something must be the case and deliberative *acceptances* (Cohen 1992), or between true beliefs and *aliefs,* which are enactive states of habits and automatic reactions in which some elements of belief are implemented along with some elements of directive states (Gendler 2008). Both distinctions aim to discriminate on the level of agentive analysis, however. Representations as they are conceived in cognitive science (e.g. Shea 2018) refer to a more primitive unit in cognition. Aliefs, beliefs, and acceptances may be considered qualitatively different constitutive products of representative mind. But if there is some truth to the enactivist story in some cases, where the behaviour implements belief-like states without cognitive representations, alief also covers these cases. The taxonomy of doxastic states is not important here. The important point is that the agent may have belief-like states attributable to her as beliefs in an agentive-level attribution without their being directly reducible to cognitive representations. This depends on how cognition works – what explains the belief-like states of an agent in the control system of behaviour. The answer to this is not important here, but the question itself highlights the difference between the different levels of conceptualizing behaviour.

Depression has effects on individuals that make it difficult to rationalize their behaviour. The origin histories of depression cannot be fully understood in terms of folk psychology, either – that is, we cannot always give a rationalizing reason for being depressed, and it may be dangerously misleading when we try. Depression simply is not a reason-like state nor a collection of reasons or desires, and neither explaining depression nor explaining *with* depression is a rationalizing explanation (cf. Goldie 2007).[135] But the difficulties surface even in specific actions in a more careful analysis.

Recall Amos and Beatrice again. Amos is drowning. Beatrice sees this, jumps into water, and rescues Amos. Beatrice has no other external goals for the action than to save Amos, such as appearing heroic, receiving gratitude, promoting romantic feelings, or signalling virtue. The reason for Beatrice's action is the combination of her desire to help him and her belief that she could do so by pulling him to shore. Furthermore, this is what Beatrice intended to do. Let us still suppose that an egoistic motivation theory of helping behaviour is true and the internal motivation structure is such that the causal process of perceiving Amos's distress and jumping to help him goes through the distress that Beatrice feels when she sees Amos in danger. This is in contrast

[135] This is not to say that our folk-psychological practices do not *influence* depression episodes. A person's beliefs about his or her own states may, for example, cause loops in which negative self-evaluations and self-blame deepen the depression or even cause it to become chronic (Bentall 2003; Beck 2008). But this is also a case of folk-psychological misattribution of causes on behalf of the person themself. In fact, conceptualizing depression episodes in terms of reasons may be highly misleading for both understanding and curing depression. People commonly make remarks such as "what reason has he to be depressed?" or believe that people could think themselves out of depression – which sounds like they are guided by folk psychology in their evaluation and solutions. More generally, psychiatric explanations such as childhood traumas should not be understood as reasons to act in certain ways, but as causes of psychological tendencies and pathologies. (See Murphy 2006 for explanation and classification in psychiatry; see also Glennan 2015.)

with the empathetic concern theory, according to which seeing *Amos's* distress is the motivating (triggering) factor for Beatrice, and her own distress is not needed. (See Batson 1991 & 2011). Even if the selfish theory were right as a scientific psychological theory, it would not make sense to say that the *reason* for Beatrice's action lies in her wanting to get rid of her own distress, as caused by Amos's plight. This would only mean that the psychological mechanism involved in altruistic actions has a hedonistic motivational structure on a much more primitive level than the level on which we specify goals of action. Let us suppose that Amos is a small child and Beatrice is her mother, and her psychological disposition to save him is both extremely robust and strong enough to even become self-sacrificing. It would be absurd to say, within the folk-psychological framework, that this is a selfish act, even if the psychological process goes through an extreme distress triggered by the situation.[136] In fact, the extreme distress could be interpreted as a source for a strong agentive-level disposition to act altruistically.

At the same time, this cannot be turned into an argument for psychological altruism proper. Regardless of the agentive-level description, there is an empirical difference between altruistic and egoistic motivation theories. Different aspects of the internal functioning of the motivation mechanisms could be manipulated to prevent the motivation to help from arising if different theories are true. Joseph Butler (1726) famously argued against psychological hedonism by showing that internal hedonistic states are not the proper goals of actions. As Elliot Sober (1992c; Sober & Wilson 1998) has pointed out, this argument confuses the different kinds of motivational factors in action. To

[136] A mother's care for her offspring is probably the original context for the evolution of empathetic core mechanism (in de Waal's "Russian doll" model of the layered structure of empathy mechnisms; see Churchland 2011), which has then become a more general mechanism. The distress in the situation is more likely to be caused by an empathetic concern than the other way around. But for the sake of argument, to distinguish two levels of description, we will concern ourselves with a hedonistic mechanism for now.

slightly paraphrase Sober, the distinction between internal mechanisms and external goals is not maintained.[137] Stephen Stich (2007) in turn has argued that Sober's argument for the existence of genuine psychological altruism (Sober & Wilson 1998) fails to make sufficiently sophisticated distinctions in the psychological architecture. I would – and will – argue that whether this argument goes through, it is a further reason to make the distinctions between different concepts of altruism and different levels of action explanation. I will return to the issue of altruism in detail in the next chapter. The point for now is not, however, simply that we might need at least two concepts of altruism to sort out the motivation behind helping. The agentive-level action explanations (she wanted to save him) and psychological-level causal explanations (the observation triggered distress that caused the action), as well as neurophysiological explanation (whatever processes are taking place on this level), are different. The agentive description is about the goals of action, not the inner psychological workings that guide it. The confusion over how to conceptualize situations like this is caused by conflating different levels.

Moreover, when we refer to behavioural traits, such as the tendency of parents to help their children in danger, we are not talking about a goal of an action in the first place, or action-guiding desires. We may say that parents have a *preference* to help their children in danger, but the concept of preference is an abstraction of whatever choices an agent is disposed to make. The causes and constitutive processes of having a preference may be anything – including social situations and structures. It is too abstract and vague a concept to understand the goal-directedness of action on the psychological level. Having a behavioural tendency to help one's children in danger is having a tendency to form the right kind of pro-attitudes in situations with children in danger – it is not having a pro-attitude to help children in

[137] Philip Kitcher (1997) seems to make the same mistake. I will discuss this later.

danger as an abstract goal, nor is there a second-order desire of this kind.[138]

Another argument to divorce the psychological and agentive comes from looking at animal cognition, situated agency, and the evolution of representational mind, as already discussed to some extent. Simple organisms behave purposefully in their environment without a representation manipulation system. The decoupling of representations from their immediate triggers comes from the need to represent the robust features of a complex world, and the metarepresentational states, such as desire-like metarepresentations of one's own needs, are probably even later than that, for planning purposes (and possibly preceded by attributing similar states to others). It is reasonable to think that human cognition is a mix of the "primitive" features and more complex cognitive processes, whether its operations are best understood as representation manipulation or something else. (See Godfrey-Smith 1996 & 2006; Bogdan 1997; Sterelny 1998, 1999, 2001b, 2003 & 2015; Hutto & Myin 2017.) It also makes sense to study animal psychologies (with drives and representations) even when we do not consider them intentional agents, and the human mind is in a continuum with them and similar to them in parts. Humans become intentional agents when they can be attributed with propositional attitudes, and this involves the emergence of the capacity for rational thinking. But this does not mean that the functioning of mind switches completely from non-intentional operating to exclusively rational processing.

The functioning of the mind and the mechanics of thinking are not important for this dissertation as such. How to explain behaviour is. But they are connected. In the next sub-chapter, I will discuss the relations between *rationalizing* action with agentive states, *causally explaining* the external behaviour through physical (neurophysiological) states and thinking. I start with the basic functionalist idea that the states of mind that psychologists refer to are on an intermediate layer

[138] Parents may, of course, *also* have second-order desires about the desirability of helping one's children in danger, but this is a different thing.

between the two – a level that describes causally effective brain activity, but as functional, computational, or algorithmic descriptions of *what* the brain processes do (see Putnam 1975; Fodor 1975, 1981 & 1990; Marr 1982; Shea 2018). This is conceptually different from the agentive descriptions at minimum, and the stronger argument can be made that rationalizations do not refer to this level – directly at least. This is also the level of *explananda* for evolutionary psychology. This discussion is a detour from the main argument, but there are complications that make a simple distinction between levels proposed here an insufficient characterization of action explanation (for example, the apparent causal relevance of some rationalizations and the role that folk psychology as a practice plays in social behaviour), so I will discuss them shortly to block some possible objections. I will not discuss the ontological or theoretical relations between the various levels of description further – the only assumption I will make is that the agentive descriptions and psychological descriptions are *conceptually separate* and do not *actually* go hand in hand *always*.

5.3. On Action Explanations

Many philosophers have argued that folk-psychological ascriptions of intentional states to agents should be analysed as functional states of a person that make action intelligible *instead* of referring to the causal structure of mind, even when they disagree on whether the states *also* capture something causal (see Davidson 1963 & 1970; von Wright 1971 & 2001; Lewis 1972; Dennett 1987; Bennett 1991; Bogdan 1993 & 1997; Nichols & Stich 2003; Godfrey-Smith 2005 & 2008; Sehon 2005; Hutto 2008; O'Brien 2019). The foundational level of causality is usually considered to be the physical/neurobiological level, and causalists connect this level to intentional states either by analysing the relationship between mental states as person-state ascriptions and physical level directly, or by treating an intermediate psychological level as being constituted by physical processes but realizing intentional states. As

stated above, these are rivalling theories of what intentionality is, but I suggest that there are two different levels of description, agentive and psychological, the former being the level of person-level attributions and the latter being the level of sub-personal causal processes. I have also suggested that folk psychology itself, as a natural practice, does not make the distinction: there is no difference between a belief as a mental representation (that takes part in cognitive processes) and a belief as a state of an agent (as its relation to the world), or desire as an agentive state and as a psychological directive state. Nor is there any reason why folk psychology as a practice should make such distinctions. For practical purposes, the distinction between states that articulate behavioural dispositions of a person and states that are causally effective psychological factors which entail such dispositions is not that important if there is a high enough covariance between them.

For scientific and philosophical purposes, different frameworks should be distinguished conceptually, whatever their relation turns out to be. Without this revision, it looks like rationalizations of behaviour causally explain it, which seems to be both true (we *do* explain people's behaviour using reasons as if they were causes) and false (agentive descriptions do not refer to causal processes). This is the classical problem of *mental causation* here, which I will not attempt to solve here. I will only argue that the (rationalizing) agentive and (causally explanatory) psychological levels of description are distinct but connected in various ways (for example, there is causally efficient rational deliberation that uses folk-psychological categories in reflection, and individual rationalizations have causal presuppositions), and folk psychology as a practice does not make the distinction. Consequently, folk-psychological explanations cannot be used directly in more sophisticated action explanation – philosophical, psychological, or evolutionary.

5.3.1. Rationality and Rationalization

An essential source for difficulties and connectedness between psychological and agentive levels is rationality. The agentive description attribute reasons to agents, and the relationship between reasons to each other and the action is rational. The rationality of action like this seems to presuppose some sort of rationality in the causal processes that produce behaviour if agentive descriptions are given a causal explanatory role – not rationality itself to be a causal factor, but the causal processes to have systematicity in their functioning that exhibits behaviour that we perceive as rational. Furthermore, rational deliberation about the goals and means to achieve them is a part of human psychology, not just a property of action attributions.

There are, however, different concepts of rationality that may be applied to action and should not be conflated, especially if we are interested in their connection to causal explanation of behaviour. **Agentive rationality** is the notion of rationality used in the philosophical theory of action: the idea that there is a reason for action. The action is rational when it is in accordance with the goals and beliefs of the agent in a way that can be expressed as giving the action a reason. There are both descriptive and normative elements in this: the action can be described as intentional by giving it a rationalizing conceptualization, but rationality is also evaluative in the sense that we consider action itself appropriate or not, given the reasons it was taken (see McGeer 2007; O'Brien 2019). Agents may be more or less rational in the sense that the action may be more or less appropriate, but *it is a qualitative property of an action that it can be given a reason*. It is about the intelligibility of behaviour as the action of an agent. The proposition that humans are rational agents is a categorical proposition about rationalization, both in its descriptive and normative dimensions. If humans are rational in this sense, rationalization is an adequate way to conceptualize humans and human behaviour. The normativity of rationality in this sense is what makes human action rational or *irrational*, while most animals, for example, are not rational or irrational but

arational (see Hurley & Nudds 2006). **Cognitive rationality**, which is the notion used in cognitive science, is a quantitative measure of cognitive capacity – but it is, likewise, also a normative notion. Rationality is measured against the *chosen optimality model,* which specifies what counts as rational, either in the *epistemic* (belief-formation) or *instrumental* (decisions about which course of action to take given the context) sense, and the *degree* of rationality and irrationality in human action and thinking is evaluated by comparing the performance to the model (see Stanovich 2011 & 2012).

The two senses of rationality, and the notions of normativity accompanying them, are different. The philosophical analysis of folk psychology uses the agentive notion. It is supposed to capture something that is *constitutive* of agency. Its normativity is about the adequacy of action given its reasons, and the failure to be rational is a failure to be an agent and for the behaviour to be intelligible as human action (see O'Brien 2019). Cognitive rationality and its normativity are instrumental: there are models that we *choose* to represent optimal decision in a context, *given the aims of the agent*, and we compare the behaviour to this. Moreover, these models (and the concept of rationality) could be applied to non-intentional systems, too, such as those animals that we consider not to be intentional, and to Artificial Intelligence systems. Agentive rationality does not imply any particular model of cognitive rationality. Moreover, it cannot be assumed that agentive rationality implies any particular *degree* of cognitive rationality that would enable *agentive rationality itself* to be an explanatory factor for behaviour, for example (see also Henderson 1993; Ylikoski & Kuorikoski 2016).

The two notions of rationality are also related. The models of cognitive rationality are *meant* to be about what a rational agent would ultimately choose, given their goals. This implies a third notion of rationality, ***normative* rationality**: how one *should* reason and choose action, given the goals. This is a stronger notion of rationality than the one used in rationalization of action – the assumption (rational) agency is not an assumption complete rationality. However, although

this is a more demanding normative notion of rationality than the other two concepts in their normative component, the normativity of normative rationality is *instrumental:* it depends on chosen goals and acknowledged constraints on achieving these goals. This is the notion of rationality for fields such as Decision Theory and does not concern us here. However, the ability to be a rational agent in the agentive sense requires some cognitive capacities that explain it. Cognitive rationality is a measurement of how well some of the cognitive capacities function in certain tasks, and having these capacities is a partial explanation for why humans are agentively rational. These capacities are what we should be interested in when causally explaining human behaviour and how it fits with the causal understanding of human behaviour that humans are also (agentively) rational. I will refer to the agentive notion as "rationality" from now on, unless otherwise specified.

An essential feature of folk-psychological explanations is that they rationalize the behaviour into actions that have reasons behind them and goals to look forward to. Reasons (and their constituents, beliefs, and desires, the "two directions of fit," as Elizabeth Anscombe (1967) put it, descriptive and directive) are connected to each other and to the action in rational relations: the propositional contents entail other propositional contents and are attributed to agents as holistic sets. At the same time, the attributions identify what action the behaviour is, and the identification of the behaviour as doing x is a part of the interpretation of which beliefs and pro-attitudes of the agent constitute their reason for action in the situation. That is, the action descriptions are a part of the same holistic net of semantic connections as the mental states that make behaviour intelligible. For some this means that rationalizations cannot be causal explanations, since semantic entailments are not causal relations (Anscombe 1967; Taylor 1966; von Wright 1971; Sehon 1997 & 2005; see also Mele 2000), whereas others think that this merely makes the ontology of action somewhat anomalous (Davidson 1970). It seems that folk-psychological practices require the attributions to have at least some causal counterfactual power: the point of persuasion and reasoning with a

person, for example, is to change their underlying structure of desires and beliefs to affect their future behaviour. This is a causal intervention, not a matter of interpretation after the fact. Mental attributions should not be causal attributions, under some conceptual and metaphysical considerations, but they seem to function as if they were. Hence the attempts to reduce the rationalizing elements into something that also has psychological reality. (See Heil & Mele 1993; Henderson 1993; Crane 1995; Mele 2000; Kokkonen 2011 & 2012.)

Furthermore, folk-psychological practices seem to presuppose that of all the reasons we can attribute to the agent, there is a *primary* reason that is why the agent actually did what they did. It determines what the action was about – it is not just an alternative description for the behaviour. How should we understand this? For a causalist like Davidson, the primary reason is the one that caused the action. Under the ascription view, this is a problem known as *Davidson's Challenge*: a mere ascription is only about pattern fitting, it does not explain action (see Davidson 1963; Mele 2000; O'Brien 2019). For the causalists the problem is how the reasons can be causes. This problem can be broken into two parts. First, how can mental states (that is, agentive states under the description of folk-psychological conceptualization) be causes of physical behaviour? I call this the *core problem of mental causation*.[139] Second, how could rationalization of action reliably capture states that are causally efficient for behaviour? In other words, how can *reasons*, identified by their modal and logical properties, be causal? I consider this the *hard problem of action explanation*.[140]

---

[139] In the traditional formulations of the problem of the mental causation, the distinction between psychological and agentive is not made – it is just about "mental" and "physical" (e.g. Davidson 1963 & 1970; Kim 1992 & 2005). In the three-level analysis, the problem of mental causation gets broken up as problems between the three levels of agentive, psychological, and physical. The "physical", too, consist of two levels, physical proper and neurophysiological, but the relationship between the two is a problem of a different kind.

[140] The problem of mental causation has several dimensions. In addition to what I call the "core problem" of mental causation, which has to do with

5.3.2. *The Hard Problem of Action Explanation*

The recently popular solution to the core problem of mental causation has been to use the contrastive-counterfactual theory of causal explanation and the manipulationist theory of causation as a framework to identify causal factors (Menzies 2007; Shapiro and Sober 2007; Woodward 2008 & 2014; Raatikainen 2010). Folk-psychological descriptions make robust but imprecise claims about causal processes and behavioural dispositions of the agents on which the behaviour depends, with relevant counterfactual contrasts. These robust states are the states of the agents. There may be psychological states that implement these more or less directly and have causal relations with other psychological states and the behaviour, and similarly with neural states – but these are further issues. What matters is that the mental descriptions identify states that have intelligible contrast classes, and the difference between the explanatory state and its contrast class is a difference-maker between the explained behaviour and its contrast class. The *explananda* and *explanantia* need not be described on the same

---

causation between different levels or kinds of descriptions, and the "hard problem" of action explanation, there is also the ontological problem of how subjective, reflective states of mind can have causal power. This is similar to the rationalization problem in that in both cases a non-physical, mental attribute is supposed to have a causal relation with something physical, and the concern is that either the mental remains epiphenomenal or the physical reality is not causally closed (a violation of the *exclusion principle*; Kim 1993). In both cases, the solution to the core problem of how to understand causation gives the direction for the solution but is not enough. I will not go into the full problem of mental causation here. I will restrain from discussing the parts that have to do with explanation. Furthermore, not all mental states are conscious, including not all causally relevant mental states, unless one defines mental states as having conscious content – in which case psychological processes would be a mixture of mental and non-mental factors. It would also make some explanatory reasons non-mental. This is a possible way to conceptualize mentality, but it is a matter of semantics that probably adds to the confusion more than it clarifies any issues.

level, for as long as the framework identifies the correct dependence relation. In other words, reasons can be causes when having a reason is the adequate identification of a causal disposition.

Furthermore, the contrast classes of explanation may be different when referring to the causal process on different levels. In fact, given that we attempt to explain behaviour that is specified with a goal, a folk-psychological description (including reasons and intentions) may be a *more* adequate way of identifying the contrast class than an alternative explanation on a different level (see Raatikainen 2010). This seems to solve the causal explanatory part of the problem regardless of what the relationship between agentive states and the underlying causal processes may be.[141] Moreover, it grants autonomy to causal explanations on different levels and fits the general pluralistic approach of this dissertation. Some other issues remain untouched with this solution, however. These include problems such as the ontological relation between the objects of the different descriptions,[142] but more

[141] Tuomas Pernu (2013) has argued that this solution only makes the different levels of causal notions equally legitimate but does not provide grounds for postulating inter-level causal interactions. It is true that this solution does not give a satisfactory account of downward causation (mental causing physical), which many consider to be the standard for genuine mental causation, since the mental remains epiphenomenal otherwise (e.g. Kim 1989, 1992 & 2005). This throws us back to the issue of what causation is in the first place, and the ontology of causation, which was discussed in the previous part. Even if Pernu's argument is valid, there is no problem in *causal explanation* of physical behaviour with mental states, if a causal dependency of the right kind can be established. I agree that this is not enough to solve the entire problem of mental causation, but it solves the part of the problem that it addresses.

[142] That is, the issue of reduction and emergence, and the role of subjectivity. The psychological level of explanation, be it connected to folk psychology or not, is usually thought to be functional explanation and both distinct from and autonomous of the neurophysiological explanations (e.g. Fodor 1981; Cummins 1983; Stich 1983; Dennett 1987; Shea 2018). Gualtiero Piccinini and Carl Craver (2011) have proposed that the functions should be understood as mechanism sketches and therefore the psychological explanation is

importantly, this solution does not touch the issue about the role of rationality and rationalization itself (the hard problem): how can we discover causes of behaviour by rationalizing action (or rather: can we, and how can we justify this practice)?

The problem has two components. First, how is it possible that humans are natural beings whose behaviour is a part of the causal structure of the world but follow the dictates of rationality at the same time? (*The role of rationality in naturalism*.) Second, is rationalization a form of (causal) explanation? There are only two possible solutions to the first part: some sort of anomalous monism (Davidson 1970), or that humans are not actually as rational as rationalization practices presuppose. There are good empirical reasons to think that humans are not fully rational when it comes to *cognitive* rationality (see Kahneman, Slovic & Tversky 1982; Kahneman 2003 & 2011; Stanovich 2011 & 2012). As discussed above, the agentive notion of rationality is a different notion, but there is a substantial connection between the two notions in *explaining* rationality. If rationality itself does not have causal powers (and it follows from the naturalistic premises that it does not) there must be something in the psychology that explains this. Rational deliberation is a part of how mind works, and it has causal consequences, but the empirical research seems to imply that it plays a limited role in

---

mechanistic. The mechanistic approach generally sees the reduction and emergence as properties that come in various degrees in moving between the levels of analysis in complex systems (see Wimsatt 2007; Craver 2007), but the details matter. For example, different levels of analysis may identify the system itself in different ways (that is, what are the important causal factors, which differences between entities or processes make a difference under which description, which entities or processes form kinds together, what counts and what the analysed phenomenon itself is, etc.). Later I will distinguish between four different levels of analysis that are relevant, and a reduction all the way to physics introduces a fifth, the level of physical processes. The ontological relation between "mental" and "physical" goes through these levels and the moves between them are likely to be different in kind. But this issue as such is not important here.

cognition. This also implies limitations on the extent of agentive rationality in humans. (See also Henderson 1993; Ylikoski & Kuorikoski 2016.) Partial rationality (whether it is because of deliberation or something else) may be enough to justify the interpretative practices, however, and it is not an unsolvable problem for a naturalistic view of humans. Humans have complex cognitive systems adapted to survive flexibly in complex, changing environments. A part of this process has been the decoupling of representations from what is immediate, and this has also created a need to represent states of affairs as related to each other and make inferences between them (Godfrey-Smith 1996; Sterelny 1999 & 2003). In other words, humans have evolved psychological processes that are causal (and implemented by neural processes) but deal with representational states in a way that is partially rational, since the psychological mechanisms have been selected for having rational outcomes. But this rationality is relative to selected tasks and their proper contexts. There is no selection for universal rationality. And even if there was, an organ such as the brain could not produce universal rationality through causal operations. Then again, we are not universally rational. This solution also makes the rationality of human behaviour and psychology (to the extent that it is rational) an *explanandum* itself – *rationality* (the entailments between propositional attitudes) does not explain rationality of behaviour. *Rationality is a part of descriptions of the behaviour to be explained*.[143]

Folk psychology is, however, also a crucial part of our social practices and the cognitive skills related to them, and it evolved to be functional to the many different needs of our many kinds of social interaction (Byrne & Whiten 1988; Whiten & Byrne 1997; Bogdan 1997; Corbalis & Lea 1999; Tomasello 2009; Emery 2012; Devaine *et al* 2014). This

[143] A part of the solution will probably be a deliberate, conscious rational thinking as *a part* of overall cognitive processing, but it is not doing the thinking alone and it is not responsible for the rationality of behaviour alone. More about this later. Deliberation is not, however, a single capacity either, but its workings need a constitutive explanation by a causal system that cannot be perfectly rational in its functioning, for metaphysical reasons.

evolution does not need to have been a passive selection of our mindreading capacities to human behaviour, either – there are equal selection pressures in the other direction. The need for effective mindreading for various social activities to be possible, entails selection pressures on our behavioural tendencies, as well as the "control structures" in our cognition, to be more in accordance with the kind of rationality that we use as a guide in folk psychology (Sterelny 2015). Furthermore, the folk-psychological practices, the language related to them (see Gopnik & Meltzoff 1997; Zawidzki 2013), and the agent-based narrative structure we learn in childhood (see Hutto 2008) affect our thinking. They do have not only mindreading but also *mindshaping* functions – they are an extra-genetic form of inheritance to shape our behaviour and its underlying psychology to be in line with folk-psychological assumptions, as suggested by Matteo Mameli (2001) (see also Zawidzki 2013; Sterelny 2015). Moreover, folk psychology has regulative and justificatory functions in social interaction (Andrews 2015a & 2015b; see also McGeer 2007; Zawidzki 2013). All this makes rationality understandable from a naturalistic point of view as far as it is limited, but rationality as such does not play an explanatory role in why we think and act rationally. The (evolutionary) function of folk psychology is to enable us to track, predict and guide patterns of behavioural tendencies, not to describe the cognitive architecture behind it. The feedback loop from the expectations of and selection for rationality that pushes us to be rational is caused by the expectations and selection, not by rationality itself. This also means that rationalization is not a mere projection of agentive states – it is a part of the causal story. But it is a part of the causal story *because* it is used in social practices, not because it captures something that would exist without these practices.

This still leaves us with the second problem, rationalization *as* explanation. So far, I have been arguing that rationality of human action is an *explanandum*, not an *explanans*. How could rationalizing with a reason itself be an explanation? One possible solution would be to revise the non-causalist stance on attribution of mental states by

proposing that psychological states (in the narrow sense) are references to causal states, but rationalizations are about agentive states. I have already alluded to something like this as the first approximation. But making this distinction alone would cut the connection between rationalizations and causal explanations, and rationalizing attributions seem to work as attributions of causal factors. We could go even further: we rationalize action only because it captures something causal. Furthermore, it would leave us with Davidson's Challenge: the notion of primary reasons, the intended reasons for action, require some further explanation if intention is not causally effective (see Mele 2000; O'Brien 2019).

5.3.3. A Causal Presuppositionalist Account of Rational Action Explanation

Consider the following option. It is not the *reasons* the agent has that are manipulated in an interaction, but something that *having the reasons depends on* (that is, something causal that can be described on the psychological and/or neurophysiological level). If the connection between the reasons and their underlying conditions is sufficiently robust, reasons identify causal relations, albeit under an imprecise description. Reasons are attributed by rationalization, and they depend on psychological processes. This would be a form of anomalous monism that is not anomalous, given there are explanations available for why the two are correlated. However, this is not a sufficient solution. Describing intentional action involves ascribing an *intention* to the agent, not just rationalizing reasons for action that can be interpreted for the agent: some reasons express what the agent intends to do, and these intention references are clearly meant to capture something causal (Davidson 1978; Bratman 1987; Mele 1992 & 2009). And as Elizabeth Anscombe (1967) (albeit a non-causalist herself) pointed out already, we also seem to have direct knowledge of our own intentions. Our knowledge of all the factors that play roles in why we do what

we do may be fallible, but the experience of intending to do something in particular is direct, not a process of interpretation. Within the causal interpretation, we identify some of our reasons as causes. Furthermore, we do not just act and interpret the action; we reason about our goals and the means to achieve them, and this reasoning seems to make some causal contribution to producing behaviour. Hence, the connection between agentive rationalizations and causal psychological processes seems to hold.

It is, however, one thing to say that we have more intuitive understanding of ourselves as agents than a mere interpretation and another thing to say that this understanding involves direct observations of the causal processes that guide our behaviour. I have already mentioned philosophical and psychological doubts for the latter. More importantly, we are only conscious of a part of our cognitive processes and motivations for action. Cognitive and social psychologists distinguish two kinds of processes in mind[144] (the so-called *dual process* and *dual system* theories of cognition): Type I (or System 1) and Type II (or System 2). Type I processes are *automatic*; they are fast, reactive, non-conscious, associative, heuristic, and effortless. Type II processes are *analytic*; they are slow and effortful but controlled and deliberative. (See Kahneman, Slovic & Tversky 1982; Epstein 1994; Evans and Over 1996; Bargh & Chartland 1999; Stanovich 1999 & 2011; Stanovich & West 2000; Kahneman 2003 & 2011; Frankish 2004; Lieberman 2007;

[144] The division into two kinds only is problematic, and the different theorists highlight different differences between the different kinds of processes (see Stanovich 2011), which may imply a much messier image of cognition. Emotions, for one, participate in much of our choices of action and they are something that we are very conscious of, but they remain mostly outside rational processing (Griffiths 1997; Haidt 2001; Prinz 2004) and their cognitive contents are probably partly non-conscious (see Winkielman & Berridge 2004; Gawronski *et al* 2006). But the important point here is only that there are both conscious deliberative processes that may involve rational reasoning, and non-conscious processes that are more associative and nevertheless participate in reasoning processes.

Carruthers 2008; Frankish & Evans 2009; Evans & Stanovich 2013.) These processes (or systems) are jointly activated, and they give a rise to more complex cognitive operations, but only some processes are conscious, and we are (indirectly) aware of only some of the non-conscious processes. We have no access to all the processes that influence our thinking, even our conscious thinking. When people are asked about the reasons for their actions, they do not identify an *effective motivation* behind them, but describe a *state with a goal*, and this may be just as much a rationalization after the fact as if they were explaining another person's action, even in highly deliberative contexts such as making a moral judgment (Haidt 2001; see also Nisbett & Wilson 1977; Nisbett & Ross 1980; Bargh & Chartrand 1999).

As mentioned earlier, the notion of rationality used in cognitive science is different from the one used in the analysis of folk-psychological conceptualizations, although there are substantial connections.[145] There are two properties of the two-level cognitive system that are consequential for the issue at hand. First, the analytic processes that we are conscious of and constitute our deliberation are the ones we identify as our thinking and decision-making in our cognitive phenomenology. We experience other states too, such as emotions, and we are usually aware that we have other psychological motivating factors, but reasoning is what we consider to be our thinking and responsible for decision-making. We can disregard the normative,

[145] Furthermore, there are two schools regarding how cognitive rationality should be understood, regarding the two levels of processes: the *heuristics and biases school* and the *cognitive ecology*. The representatives of the heuristics and biases school think that there is a great deal of irrationality in human thinking, with the analytic processes being the rationalizing part (Kahneman, Slovic & Tversky 1982; Kahneman 2003 & 2011; Stanovich 2011 & 2012). The cognitive ecology school thinks that the simple automatic processes guide human behaviour to be more rational in an adequate context even if the processes producing this outcome do not work in rationalizing ways (for example, Anderson 1990; Cosmides & Tooby 1996; Gigerenzer 2007; Marewski, Gaissmaier & Gigerenzer 2010). Whoever is right about this does not matter here.

gradual notions of cognitive rationality for a while and concentrate on some of the qualitative aspects of the analytic processes. First, they process propositional contents: this part of cognition is closest to what folk-psychological rationalization presumes human thinking to be like. Second, our thinking and decision-making involves the non-conscious processes as well, even while we deliberate, and they have inputs into the deliberation. When we deliberate, we become aware of the products of non-conscious processes as our own thoughts (even if we do not have access to the processes producing them), and they become a part of further deliberation. Third, the agentive rationalizing attributions to agents (as whole persons) do not distinguish between these two kinds of processes.

If folk psychology is pluralistic both in its mechanisms but also in its reference, this extends to self-reflection. When we reflect our motives and decisions, we attribute the rationalization of intentional states (desires, beliefs, reasons) to ourselves according to folk psychology. The source for these states includes both the deliberative process and the other processes that participate in guiding our thinking and behaviour. If this is the case, the object of reflection on our own mental states is a combination of deliberative conscious states, products of non-conscious processes that we are aware of and interpret in folk-psychological categories, and quasi-theoretical assumptions about ourselves that are folk-psychological postulates. Our self-understanding is fallible regarding these differences. Even if our self-attributions of mental states are correct in terms of folk psychology in the moment of action, and even if they are based on epistemically reliable self-observations, our *justificatory* self-rationalization does not necessarily identify the *causal* processes of how we came to the decision correctly. Furthermore, we are not necessarily correct in our self-attributions either, and our self-observations are not always reliable.

However, reflection is not mere rationalization. Sometimes we explain our own behaviour with non-rational causes, such as anger, sorrow, or intoxication. But the point here is that sometimes we also misidentify having non-rationalizable psychological processes as

having reasons. Conscious reasoning (as a part of cognition) and interpretation using the theory of mind (on the agentive level) are confused in the simplified image of rational agency, and they should be distinguished. Moreover, although we experience intending and identify it correctly as the motivational state that triggers action, the content (the reason, or a plan) we accompany it with may sometimes still be an interpretation within a folk-psychological conceptual framework, not an experience of a deliberated state. There are also problems with prediction of one's own behaviour: people are notoriously bad at predicting their future actions based on their current self-perceived states of minds – although it is not clear whether this is because of misinterpretation of one's own motives or underestimating the situational factors that are not present in the context of prediction (Poon, Koehler & Buehler 2014).

Having an intention (in the sense of intending) does, however, presuppose that there is at least one causal factor that is identified in experiencing intending. Psychologically speaking, we experience motivational forces, aversive and incentivising saliences that guide our behaviour. A successful agentive explanation does not need to specify these processes precisely to be a form of causal explanation. But a successful agentive description must include a reference to the *existence* of such factors. The identification of an intention in the context of action involves attributing a reason that adequately describes the agent's relation to the world in a robust way in the context, given both her epistemic states and active motivational forces. For example, in the case of Beatrice saving Amos, there is an identification of an intention to help with a propositional attitude that (correctly) describes Beatrice's pro-attitude to save Amos (as an agentive state), which is adequately paired with a motivational force pushing in this direction. It is a further question whether this motivational force is Beatrice's own distress or empathetic concern – which, on the other hand, is a real difference between two causal routes to motivation. In other words, a successful agentive description correctly refers to the agent's effective states within the conceptual framework of folk psychology, but this is not the same as

identifying the agentive states with the psychological states. Moreover, even if we think the agent herself *knows* what she is doing (or what her intention is), this does not require her to know all the psychological processes involved. On the other hand, when Beatrice reflects on her own motives, or an observer of the situation wonders whether Beatrice is, indeed, altruistic or has some other goals in mind, the observation or speculation (depending on who is doing it) targets the psychological states, but this may still use the same goal-directive semantics.

Rational deliberation is a *part* of our cognitive capacities. It is a part of the causal makeup of mind, not just a passive reflection of cognition. But reason, in this sense, is not a determinative factor. At the same time, the object of rationalization is the action, not a partial factor of it: we use folk psychology to represent our own holistic agentive states, such as beliefs and desires, in metacognition (whether in conscious deliberation or in automatized processing, which also has metacognitive functions). We do not represent just the reasoning part of our cognition, although this is the part we mostly identify our thinking with, and we tend to conflate the two – the contents of reasoning and the holistic states. Folk-psychological categorizations affect how we deliberately plan our actions, but once again, this is causal influence of folk psychology on our cognition – it does not make agentive and psychological states the same. The same applies the other way around; not all behaviour needs to be produced by deliberation alone in order to be rationalizable in the sense of agentive rationality. Much of the unconscious, automatized processing has a positive function in reaching the chosen goal of action (see Bargh & Chartrand 1999; Gigerenzer 2007; Marewski, Gaissmaier & Gigerenzer 2010).

For example, if I want to get some coffee from the department's break room while thinking hard about how to formulate an example of automatized processing for my dissertation, and I have made this same trip literally hundreds of times, the only conscious choice I make might be the decision to get coffee, while all the guidance of the rest of the action might be automatized, including opening doors and pouring coffee. This involves automatized observations, inferences

and behavioural decisions guided by a learned script and almost no conscious thinking at all. Similarly, automatized processes may guide behavioural choices and decision-making even in much more complicated situations. Developing routines is a part of the skillset of any profession and much of our evolved psychological capacities work like this rather than in a conscious way. If the agent has chosen a goal and starts acting towards it, if the automatized processes work towards this goal, and if the agent identifies with the behaviour they are participating in the production of, this is sufficient for it to be an example of agentive rationality.[146]

We can summarize the discussion above in the following propositions about rationality and action: (1) Reasoning (as rational deliberation), is a cognitive process that participates in the causal production of behaviour. (2) People are conscious of this part of their own cognitive processes, while the other processes manifest only in the products of these processes. (3) The non-conscious parts of cognition are often instrumental to the chosen goals, and therefore they participate in producing the action that we rationalize *without being a part of the rational guidance* of the action on the cognitive level. (4) People rationalize both their own and others' actions within the folk-psychological

[146] As inferences become automatized with practice, the agent might lose the sense in how they do what they do and may give completely false interpretations of themselves, as discussed here. A rather striking example of this phenomenon, although anecdotal, is when a magician or a charlatan fortune-teller doing "cold readings" starts believing themselves to have supernatural powers. "Cold reading" is the art of gathering information from a person and feeding it back to them, making guesses and testing them, and giving an impression to the person that the source of this information is supernatural. A cold reader may become so good at this that they lose the sense of how they do it and start to experience their own skills as something supernatural. According to Orson Welles, who was a magician, among other things, and learned cold reading from old-time charlatan fortune-tellers in the 1920s, this was a well-known "occupational disease" among the charlatans who themselves called this phenomenon "becoming a shut eye" (an interview on The David Frost Show, May 12, 1970).

framework and this rationalization has more to do with justification and evaluation than causal explanation, but it functions both ways. (5) People conflate the *rationalization* they apply to their action and the *experienced intention* that triggers this action *whether it is the outcome of rational deliberation or some other process*. It can be either. To the degree that non-rational processes are instrumental to chosen goals, this does not make a difference in understanding the action as guided by reasons. But giving only reasons in the causal explanation misidentifies the causes. (6) People are aware of non-rationalizable causes such as emotions and use them in the folk-psychological explanations as well, and these explanations are not rationalizations. Emotions, personality characteristics, reasons and other factors are not separated as being on different levels; but there is only one folk-psychological "level". (7) Sometimes, people have no idea what motivated them, and their rationalization of their own action is simply incorrect.

If these conclusions are accepted, agentive, rationalizing descriptions do not refer directly to psychological processes with causal powers, but they *presuppose* that there are causal processes that are responsible for the action in order for the folk-psychological practices to work.[147] In these practices, agentive attributions of reasons and attributions of psychological states proper that underlie the agentive states are mixed into a heterogeneous category, and the distinction between them would not make a difference. The connection between agentive ascriptions and the underlying psychology is strong enough to allow rational arguments, persuasion and other folk psychology-based practices to enter the cognitive system and influence behaviour. In philosophical and scientific scrutiny, however, different levels of description need to be acknowledged. Slices of the causal process that result in behaviour can be described on any level, although they do not make

[147] Lilian O'Brien (2019) has argued for a similar presuppositionalist account within action-theoretical framework. She argues that rationalizing action explanations presuppose underlying causal factors, but causality does not play an explanatory role in rationalization.

the same causal explanatory claims (since they have different contrast classes) and sometimes there is no rationalizing action explanation at all – that is, when the behaviour under scrutiny is irrational. For psychological and philosophical purposes, however, the two levels should be kept apart.

This also means that the action-theoretical framework in philosophy is revisionist and not psychologically realistic. To clarify, this is not an argument against philosophical action theory. Since folk psychology is a part of the practical view of human agency, fields like decision theory and moral philosophy[148] require a conceptual scheme built upon its core framework. Philosophical psychology and action theory are conceptual clarification *and* revision of folk psychology for analytical purposes. These are normative projects. For example, the issues of free will and moral responsibility are important substantial issues that require an adequate, precise philosophical theory, even (or rather especially) if the aim is to analyse what scientific research does and does not say about these matters (see Mele 2009). Philosophical psychology cannot be completely divorced from assumptions on the human mind, but it idealizes human thought and action in terms of rationality on the agentive level, for the purposes mentioned above.[149]

---

[148] It may be that moral judgments shape our folk-psychological interpretations rather than simply use them (Knobe 2007), and if so, philosophical psychology (as an elaboration of folk psychology) is not only a part of metaethics (as philosophical moral psychology), but an integral part of normative moral philosophy.

[149] A similar point can be made about the philosophical literature on collective action from the point of view of action theory and philosophical psychology (for example, Tuomela & Miller 1988; Tuomela 1995 & 2007; Gilbert 1996; Bratman 1999; Miller 2001; List & Pettit 2011; Ludwig 2016 & 2017). Social action could be defined as action that takes place in a social context (taking other agents into account in one way or another), whereas collective action is a special type of social action where a group of people have a common goal and work as one. According to the philosophical work, conceptualizing this within the framework of philosophical psychology requires collective intentionality of some sort: joint attention, shared beliefs, collective acceptance, we-mode intentions, or something along those lines. One issue in the discussion is whether this

It may also be that scientific psychology cannot be divorced from folk psychological categories, if the description of the phenomena and understanding the results requires it to be applied in real-world situations guided by folk psychology. This is especially true in practical branches of the scientific study of mind, such as psychiatry (see Murphy 2006), but these complications also emerge when we try to understand social behaviour – the agentive descriptions are part of how people perceive each other in these contexts. The scientific understanding of behaviour is, however, another matter.

I will concentrate on causal explanation of behaviour from now on. The main point of the discussion in the past two sub-chapters has been to show that folk-psychology and folk-psychological intuitions are not a good guide to understand human behaviour from scientific point of view. Instead, causally relevant psychology and agentive-level ascriptions should be distinguished, even if folk psychology is not good in doing so, and the two are influencing each other in complex ways. Identifying behavioural traits should not rely on agentive ascriptions, either. This does not make agentive level irrelevant, but the aims of philosophical action theory, for instance, which clarifies

---

collective intentionality can be reduced to individual intentionality or not. This is closely related to collective responsibility and the responsibility of individuals who take part in collective action (see Miller & Mäkelä 2005; Miller 2006), a topic that is important to get right from the moral and political point of view. In contrast, scientific attempts to make sense of joint action (see Tomasello 2009 and Suchak *et al* 2016 for biological models; see also Sterelny *et al* 2013) attempt to discover what capacities animals or humans need to have to collaborate, how these capacities are used, and what their evolutionary origins and the dynamics of their evolution might be. The interpretation of action by analysing it with intentional states involved and the psychological capacities have different aims here. However, the disconnect is not quite as straightforward as this. Social behaviour involves the participants using folk psychology – it is a part of the phenomenon, if not of explanation. The characterization of the capacities in terms of theory of mind may require concepts that refer to collective intentionality (see Tomasello 2009; Hakli, Miller & Tuomela 2010; Gallotti & Frith 2013). Nevertheless, this is not a central issue for the work at hand.

the ways we should understand the agentive level ascriptions, are different from the scientific aims – which is nothing controversial in the context of action theory, either. But it is important to keep the analysis of psychology, agency, and behaviour separate and not mix intuitions. This will be especially important in the next chapter, which discusses social interaction, different concepts of altruism, and reciprocal altruism as an example of an interactive trait. Psychological, behavioural, and agentive levels all play an important role, but they need to be kept apart in the analysis.

5.4. Explaining Behavioural Traits

It is now time to return to the main question of the chapter: how to identify behavioural traits as the object of explanation – this time *without* the folk-psychological framing. I will differentiate behaviour and psychological traits, now disconnected by disregarding the folk-psychological connective, as different *explananda*. Thereafter I will discuss what this means for the evolutionary explanation of such traits.

5.4.1. Behavioural Traits Revisited

Earlier I suggested that a sensible way to identify behavioural traits is to relativize them to current explanatory interests. In the most general terms, this means that a behavioural trait consists of those external behaviours that go together either functionally (being instrumental to a goal that has a specific adaptive significance in the organism's life form), by shared proximate mechanisms, by shared development, or by shared evolutionary history. Going into greater detail, there are different kinds of explanatory questions to be asked within each dimension, as discussed in the previous part of the dissertation. The various human behavioural sciences attend to various causes of behaviour in ways that are not directly combinable (Longino 2013). At the

same time, different approaches make presuppositions that may turn out to be contradictory even if the explanatory questions are separate, and comparisons can only be made case by case (Mitchell 1992 & 2002). This implies that pluralism about behavioural *explananda* must be acknowledged.

From a theoretical point of view, we may think of behavioural phenomena as intersecting causal processes that we can slice into explainable parts with precise explanatory questions, on different levels and in different dimensions, with different partial explanatory factors in mind. I discussed evolutionary functionalism as one possible (and partial) perspective on how various factors are connected, and this is the perspective discussed in this dissertation. But the hierarchical simplification of how the dimensions are related that I described in the previous chapter is a metatheoretical tool of thought, not for a genuine theoretical unification, and it works only as a starting point. There is no general theoretical way to integrate the approaches of different dimensions. But perhaps we can understand their interaction and intersections in particular cases and have a pragmatically integrated view of the phenomena (cf. Mitchell 2002), and my suggestion is that evolutionary functionalism is an instrument to achieve this in some cases. In the human context, however, this is problematic. Behaviour is directed at goals that may be abstract, and we tend to conceptualize this in folk-psychological ways that conflate behaviour and its proximate psychological mechanisms into action. All explanation of human behaviour inevitably starts with using folk-psychological concepts in categorizing the *explandum* itself (see Longino 2013).

I defined behaviour earlier, following Fred Dretske (1988), as the activity of a system, such as the bodily movements of an animal or a human being. I defined behavioural traits as categories that capture all the behaviour that attempts to achieve the same thing in some sense, in reaction to the same types of situations. Categorizing behaviour as a behavioural trait requires generalizability of the reaction in a context with an idea of what the behaviour achieves in the context. This means that a behavioural trait includes more than just the

individual behaviour. The identification becomes more problematic with flexible and complex responses. Rationalizations are one way to structure complex behaviour into units that can be used as a basis for categorizing behaviour into traits. However, Dretske does not think that reasons are *causes* of behaviour. He suggests that reasons structure the behaviour adequately by stating what it is for. The same activity (say, going to kitchen to get something to drink when thirsty) can be composed of different movements and neural patterns in different times and we need criteria for what counts as an instance of a particular kind of behaviour. We can identity this using folk psychology. As argued above, however, this is not a satisfactory solution for all purposes at least.

Brent Enc (1995), in contrast, defines behaviour in relation to the environment in which it takes place – the same bodily movements constitute different behaviour in different contexts, and different bodily movements may still be an instance of the same behaviour in the same context. Without going more deeply into this suggestion, both Dretske and Enc argue for the need to relate the causal processes (bodily movements, internal mechanisms, and the triggering causes) to something external that tells us what the behaviour *is about* (see also Longino 2013). Some account of what the behaviour is about is needed in understanding animal behaviour too, if we want to talk about behavioural traits instead of mere reactions. We can call this *quasi-intentionality* of behaviour. I will return to this in the next chapter. Helen Longino (2013), in turn, argues that since our understanding of behaviour seems to presuppose some sort of teleology like this, including in the context of human behavioural sciences, the study of a particular kind of causal factor may be highly misleading. No explanation is explanation of *the* behaviour. The *explanandum* is always a specific causal process and its effects on behaviour, but this cannot be identified with the teleologically understood behaviour. We cannot explain behaviour; we can only rationalize it and explain effects of various kinds on it. (See also Mitchell 1992.)

But perhaps we could have an alternative way to understand what the behaviour is about, relating it to something external, that would also help to relate the various explanatory factors? I suggested earlier that animal behaviour could be structured into traits with adaptive functions, and this might also be a starting point to relate other explanatory dimensions in some cases. This is true at least when the evolutionary approach is the chosen perspective. I also argued for evolutionary functionalism in general, which is why the evolutionary approach might be a useful approach to human social behaviour. In the context of human social behaviour, however, folk psychology complicates this in two ways: not only is folk psychology an alternative way to understand behaviour, which misleads intuitions, it is also a part of the social psychological makeup of humans and their social practices. This makes the folk-psychological categorizations themselves a constitutive part of the behaviour we seek to understand. Furthermore, reasoning is a part of human behavioural guidance system.[150] This means that folk-psychological interpretations move from the side of the *explanans* to the side of the *explandum*, but we still need an account of how they work. In other words, we need psychological, agentive, and behavioural perspectives. Furthermore, folk psychology may be relevant to understanding behavioural traits because *its categories are a part of what constitutes behavioural traits*.

[150] This also holds in evolutionary anthropology, including in understanding the cultural heritability of behavioural traits: they are partly transferred by symbolic representations and behaviour-copying that utilize folk psychology. But this latter part is about the acquisition mechanisms and the mechanisms that guide the behaviour as a part of the *explananda*. It does not make folk psychology a part of psychological or evolutionary *explanans* of behaviour as such.

5.4.2. *Evolutionary Psychology Done Properly*

If we make the distinction between agentive, psychological, and neurophysiological levels of *explanation*, and combine it with Marr's (1982) *levels of analysis* distinction of implementational (physical), algorithmic (representational), and computational (functional) levels, we get four levels of description to explain psychological phenomena. (1) At the "bottom," we have the neurophysiological level. Explanations on this level are "purely" biological and causal. (2) The next level is the description of how the cognitive system works: the level of representations, drives (salience), and processes involving them, of *how* the system accomplishes the cognitive tasks. It is an abstraction of what takes place on the neurophysiological level (and may include processes outside the brain[151]), but it is still supposed to capture the same causal processes. (3) The functional level describes the tasks executed by the system: the cognitive and behavioural tasks, or *what* the cognitive system is doing. This is the perspective on how to slice the *psychological processes* into traits, similarly to what I have said about slicing behaviour into traits. This is the level that requires a perspective on the functioning of the cognitive system and what it does in its environment.[152] (4) Finally, there is the agentive level. If we individuate the function of

[151] In Marr's distinction, this level is a description of what the brain does, but the implementation may involve much more than the brain – there are causal feedback loops between the brain and the body, as well as the surroundings. The 4e-approach to cognition mentioned before (the e's standing for "embodied", "embedded", "enactive" and "extended"; see Newen, De Bruin & Gallagher 2018; see also Clark & Chalmers 1998; Bechtel 2008b; and Hutto & Myin 2017) may be a more accurate way to approach both the cognitive processes and their implementation, but this does not change the topic at hand.

[152] The traditional cognitive science perspective that Marr takes is methodologically solipsistic in Fodor's (1980) sense: the internal states of the system are individuated in relation to other internal states only. Even if this were the right way to go, it would not mean that the individuation of the *tasks* should be solipsistic.

the system through the agentive understanding of what the agent is doing, this creates a link between folk psychology and psychology proper (the "algorithmic" level), but it is one of individuation and not identity, and it remains problematic for the reasons I have given above.

As discussed in the previous chapter, evolutionary psychology is a functional-level approach to the cognitive architecture and, as such, can be either a heuristic and minimally explanatory approach to the structure and processes of human mind (at the "algorithmic" level), or a historical explanatory project of how these capacities evolved. The latter is trivially a different explanatory dimension from any of the others discussed above. But how about the first? Let us assume that evolutionary functionalism (as defined in the previous chapter) is an exceptionally successful heuristic to discover the human cognitive architecture. This would mean that we could understand mind in a framework that makes the known cognitive processes and mechanisms sensible and understandable as solving tasks in particular cognitive ecologies, and new discoveries would be predictable. We could organize our knowledge of cognitive processes and motivational mechanisms in evolutionary terms. This is precisely the aim of much of evolutionary psychology, as discussed in the previous chapter. If evolutionary psychology is understood in this way, as I think it should, it means that these functionalist descriptions do not capture additional psychological factors (on the algorithmic level) and the functions it uses are not only *instrumental* but on a different (explanatory) dimension from the folk-psychology-driven functional analysis of psychology. Furthermore, evolutionary psychology does not provide another layer of proximate mechanistic explanation for why we do what we do, or what causes us to do what we do.[153]

[153] Confusing psychology proper, evolutionary-psychological functions, and folk-psychological functions is quite usual in the popularizations and public understanding of evolutionary-psychological research, by proponents and critics alike, and probably by some of its practitioners too. This might be a case of folk-psychological meta-theory-ladenness.

Evolutionary psychology is (or should be) a way to think about the cognitive processes and motivational forces in order to learn about them – and perhaps about their origin. In contrast, agentive-level descriptions refer to the behavioural dispositions of the whole individual, and the functional perspective on cognitive processes derived from this traces what explains these dispositions of the whole system. The evolutionary approach only traces the fitness consequences and adaptive fit to an environment of the cognitive processes. It is another dimension of explanation, not another level of description. The evolutionary "reasons" (the adaptive functions) are not reasons for action. This is almost trivial, but sometimes "evolutionary reasons" are treated like folk-psychological reasons, as if they were agentive states, the "ultimate reasons" for action – also mixing the two senses of "ultimate"[154]. This is a category mistake that confuses proximate and evolutionary *dimensions* (not levels). Folk-psychological reasons are functional descriptions of the agent's goal-directedness, whereas evolutionary reasons are either a part of the causal past of the capacities (as historical, population-level explanations), or they are evolutionarily functional descriptions of the capacities in relation to a set of types of tasks. An individual agent may or may not ever choose to accomplish the task tokens of these types, and the "evolutionary aims" of the drives may or may not ever also play a motivational role in the agent's intentional action in which these drives participate. They play no role in rationalization, either – they are not hidden reasons comparable to folk-psychological reasons. This would be a *Freudo-Darwinian fallacy* (Ylikoski & Kokkonen 2009; see also Buller 1999 & 2005), which is nevertheless a common occurrence in popular evolutionary psychology.[155]

---

[154] That is, ultimate desire (causing the consequent proximate desires) and "ultimate" simply as a reference to evolutionary reasons in Mayr's distinction between ultimate and proximate. I explained in a previous chapter why I do not use this distinction.

[155] Elliot Sober (1998c) has raised another, related concern with using agentive and deliberative analogies as a misleading "heuristic of personification" in the context of social evolution. This way of verbalizing social evolution confuses

If evolutionary psychology in any of its versions (see the previous chapter) works, the functional level of description is nevertheless its target. Folk psychology (or to be more precise: the agentive stance) and evolutionary psychology are, then, alternative (but not competing) ways to slice both the behaviour and psychological capacities and motivational mechanisms into traits. Folk psychology probably plays both evolutionary and developmental roles in shaping the mind, and its developmental role may have an evolutionary reason (Mameli 2001; Zawidzki 2013; Sterelny 2015), but this is different – in this case, the folk-psychological description is a part of the phenomenon, within the evolved practice. Consequently, people have motivational states that can be understood from an evolutionary point of view, but these should not be confused with motivational states that are understood in folk-psychological terms. Moreover, evolutionary approaches are about types of motivation mechanisms and their effects on the population level and/or during the lifetime of the individual, whereas the folk-psychological approach is about the intentions of the agent in action.

5.4.3. Evolutionary Explanations on Other Levels

Evolutionary explanations can also take place on other levels. The primary locus in the evolution of mind and behaviour is, of course, the brain (and its interaction with the body and the environment). For example, Stephanie Preston and Frans de Waal (2002; see also Preston 2007; de Waal 2009) discuss neural mechanisms and how they become layered (almost literally) in the evolution of brain while discussing their *perception-action model* for empathy and the evolution of empathy, and how this is connected to the evolution of altruistic behaviour. The level of implementation places constraints on what can be implemented, and although the functioning of mind in abstract also sets constraints on further evolutionary steps (through interconnectedness, for

---

the deliberation of possibilities and the selection of actual fitness consequences as what directs the evolution of social behaviour.

example), the brain itself is an important factor in this. Furthermore, identification of the brain areas involved in psychological mechanisms can help to characterize the operations, even if this does not directly reveal them (Bechtel 2008a; see Brüne & Brüne-Cohrs 2005 and Chakroff & Young 2014 as examples of the practice), giving an opposite direction to the task-functionality to decompose the cognitive processes into mechanisms and traits. Furthermore, the behaviour itself can similarly be sliced into traits using evolutionary functionalistic approaches. This is done in evolutionary anthropology. Moreover, given the environmental, symbolic and behaviour-copying routes of the inheritance of behaviour, the historical evolution of behaviour is not entirely dependent on the genetic evolution of mind, as discussed in the previous chapter.

The proximate (psychological) dimension of explanation has levels of description (neurophysiological, algorithmic, computational, and agentive), and there are other dimensions: evolutionary and developmental. The behaviour can be integrated into traits using an agentive description, an evolutionary functionalist description (of the behaviour or of the cognitive function), or by identifying a psychological mechanism (a type of processing) that underlies the behaviour. However, if the identification of a psychological mechanism requires functionalist criteria, this loops back to either agentive, evolutionary functional, or some other quasi-teleological characterization. This does not make the identification of the psychological mechanism identical with this characterization. Regarding agentive characterization, this would be the case only if the reasons for action that cause the tokens of behaviour to be tokens of the same activity were psychological mechanisms themselves. Regarding evolutionary functionalism, this would be the case if the psychological mechanisms were modular to the degree that no two functionally different behavioural traits would ever use the same underlying processes. Neither is the case (see Buller 2005). Instead, focusing on the evolutionary functionalist characterization, the function of the psychological mechanism is what it does that is (or was) positive for fitness in its effects in all the contexts where

the same mechanism is used.[156] We can slice the behaviour into behavioural traits based on their joint achievements, in accordance with evolutionary functionalism, as in the example of the hunting behaviour of the honey buzzard. The psychological processes bundle into traits in accordance with the type of the cognitive task achieved by the trait: for example, the more general capacities that the honey buzzard utilizes in the hunting behaviour but not only there. Even if evolutionary functionalism is the perspective on functionality in both cases, the functional decomposition is different. This means that there is no isomorphic mapping between psychological and behavioural traits from the evolutionary point of view either.

Development, as the third dimension, brings further complications. First, the object of selection is not just the trait but also the development of the trait. The dependencies in the developmental process bind psychological structures together, providing an alternative way to define either psychological or behavioural traits. This may be unintuitive (but see Bjorklund & Pellegrini 2001) and I am not going to use it as a criterion. However, the connectedness itself is important for evolutionary explanations. Furthermore, many human behavioural sciences approach behaviour from a developmental perspective (for example, studying hormonal effects, or the various genetic approaches) and although the starting point is the general (folk-psychological) understanding of a trait, these approaches may end up identifying the behavioural

[156] As discussed before, identifying a mechanism always requires an idea about what the mechanism is *for*, which makes the functionalist criterion necessary. But this is not the only criterion – since a mechanism is constituted of parts (e.g. cognitive processes), we can identify the same psychological structure as a mechanism that takes part in several activities. It is a mechanism for several things. We could also choose to say that the same processes *implement more than one mechanism* instead, but if we are interested in the structure and functioning of psychology on the algorithmic level, it makes more sense to identify this structure as one mechanistic part of the psychological architecture – even if we differentiate between the two functions on the computational level. However, this is mostly a semantic issue.

traits that they study somewhat differently from other approaches (see Longino 2013). Moreover, behavioural genetics works on the population level and molecular genetics studies individual developmental processes, which means that they have different objects. Secondly, when we acknowledge the role of development, and both the plasticity of the development of the psychological capacities and the plasticity of behaviour with identical psychological capacities, and include cultural and other social factors that affect development, folk psychology enters the picture once again – but as a mechanism that participates in the acquisition of behavioural dispositions.

5.4.4. The Scope and Specificity of Behavioural Traits

Different scientific approaches may identify a trait differently, given their specific explanatory interests and methods. But the differences are not only about different levels and dimensions – their targets may be different in their *scope* and *specificity*. The behavioural trait we are interested in may be more specific or more general: we may be interested in explaining a specific type of helping behaviour in a specific context, a wider range of helping behaviour types, or altruism in general. In addition, even after we fix the class of behaviour and its specificity, the explanation may have different scopes. First, one may be interested in explaining an *occurrence* of a behaviour – for example, a particular occurrence of helping behaviour. Folk psychology is usually concerned with this and refers to the reasons of the agent for the specific actions in specific contexts, such as an intention to help a person for no other reason than helping them. The *individual tendency* to behave in a certain way may also be a target of a folk-psychological explanation. This might involve a reference to personal characteristics (such as an altruistic personality). These explanations presuppose an underlying psychology, and a psychological explanation would refer to the existence of these cognitive capacities and motivational dispositions as the basis for such a reaction. For example, the psychological explanation for an altruistic tendency could include either an

empathetic concern or a tendency to become distressed when others are in need of help – both of which can result in the same contextual behavioural disposition. This reference may be to a species-typical psychological mechanism or to the individual dispositions of the agent. Either way, the explanation on the psychological level like this is not about the occurrence, but about the *capacity* for the behaviour, regardless of how it manifests. An anthropological explanation, in contrast to both, would refer to the context and the function of the behaviour, which is about constant habits instead of occurrences or capacities. An anthropological explanation can be used in an explanation of the occurrence, but the proper *explanandum* is the *tendency* of a member of the cultural group to behave in this way.

Another class of explanations is those for the *existence* of the behavioural trait or the psychological basis for it. The explanations for existence come in two dimensions: evolutionary and developmental. The object of a developmental explanation may be the neural system, the psychological capacities, or the external behaviour, and the explanations may concentrate on (molecular) genetic factors, the external factors, or the interaction in various ways.[157] The explanation of existence may also have different scopes: the existence of the trait at all, its dominance, or its frequency. All these have both evolutionary and developmental dimensions. This (and only this) question also brings in behavioural genetics in its peculiar way to abstract a subset of factors from individual developmental processes to explain variation on a population level.[158]

---

[157] Therefore, there are different fields that study the development of psychology and behaviour, such as behavioural genetics, developmental neurobiology, developmental neuropsychology, and developmental psychology, as well as research projects such as Developmental Systems Theory (which will be discussed more closely later) that aims at rethinking the process of development and integrating various factors. Longino (2013) provides an excellent but partial review of different approaches and how they trace different causal processes.

[158] This abstraction of a subset of factors, genes, also involves an abstract concept of gene that is not reducible to the molecular gene (see Portin 1993; Moss 2004; Griffiths & Stotz 2013).

Different explanations may have different specific objects and different scopes, which causes the explanations to be about a different trait, not just different aspects of the same trait. Moreover, the scope can be chosen first: there may be a phenomenon that we find puzzling (for example, altruistic behaviour from the evolutionary perspective) and this fixes the explanatory interest. As the research proceeds, the discovery of explanatory factors in a given dimension from a given perspective may end up making further distinctions or lumping the behaviour together with something that was not a part of the same trait in the first approximation. A dialectic like this is unavoidable, but if different approaches start with the same understanding of the *explanandum* but move in different directions, the plurality of approaches may be lost.

We may be able to integrate the multitude of approaches case by case by showing how the various levels, dimensions and scopes are related in a particular object of research (Mitchell 2003), or we may not (Longino 2013). Even in the latter case, the approaches provide partial knowledge. As Longino (2013: 150) says: "Partial knowledge is no less knowledge of being partial." When our ability to explain different aspects increases, our understanding expandsas well, even without integration. However, integration deepens our understanding. Our knowledge of how things fit together increases our understanding of the dependencies between processes and mechanisms. It may be that all attempts to integrate different partial explanations are just as much perspective-dependent and local as the explanations being integrated, but even partial integration is integration. If we wish to understand how the details of proximate and developmental mechanisms affect evolutionary explanations of behaviour, we need to have an integrative framework for this particular case.

6. Altruism and Other Forms of Social Behaviour

In this chapter, I will discuss social behaviour and *interactive behavioural traits*. Interactive traits are traits that emerge in the interaction between individuals. I suggested that this kind of trait may be the object of selection in the introduction; now I will give an argument why this is so, making the *object* of an evolutionary explanation a supra-individual trait and the resulting adaptation holistic rather than individual. Interactive traits can include behaviours such as dominance and mating, but the main interest will be in altruism, division of labour, and mutualistic cooperation. I will discuss them in general but focus on altruism, despite the recent shift of interest in the evolutionary analysis from altruism to cooperation (see Sterelny 2012; Tomasello *et al* 2012; Sterelny *et al* 2013; Forber & Smead 2015). This is for several reasons.

My main example of an interactive trait is reciprocal altruism – it is simple enough to make it an illustrative example, and very much discussed. I will discuss the various concepts of altruism in detail to build the ground for using reciprocal altruism as a locus for argument, but it will serve two other purposes as well. First, my discussion illustrates the points I argued in abstract in the previous chapter and shows why they are crucial for understanding social behaviour from the evolutionary point of view. The distinction between the form of behaviour and a psychological trait is very sharp here, as well as their differences with an agentive description, making it a suitable case for illustrative purposes. Second, altruistic behaviour has been the central phenomenon for much of the debate over individualism and holism in evolutionary explanation. It has been the main theoretical challenge for understanding the evolution of social behaviour and the main reason to introduce group selection models. I will later argue that even if we leave aside the problem of evolutionary altruism, which might make a holistic approach necessary on the *evolutionary dimension*, even individualistic evolutionary approaches may require us to understand the evolving *behavioural traits* as multi-individual interactive traits.

The various forms of mutualistic collective behaviour (direct cooperation and indirect cooperation through the division of labour) are easier to conceptualize as a group of individuals combining their efforts for joint goals and interests, which makes the interaction instrumental for everyone separately; the behaviour of others is just a contextual precondition for such action. But with altruistic behaviour, in contrast, the very aims of action are the needs or goals of others. I will also argue that explanatory models of altruism based on benefits from reciprocity over time require us to understand individual behavioural contexts as parts of interactions between multiple agents that extend over time – that is, the traits are interactionist.

I will begin by clarifying the notion of altruism. The term "altruism" refers to the tendency to be prosocial.[159] It refers to helping behaviour, as well as things like sharing and consolation, in both humans and animals, as well as the psychological tendency to behave this way. There is a well-known distinction between *psychological* and *evolutionary* altruism (Sober 1988a & 1989; Rosenberg 1992; Wilson 1992; Sober & Wilson 1998), in which psychological altruism is defined in terms of what motivates the action, and evolutionary altruism in terms of fitness consequences. Reasons to distinguish at least a third notion of altruism, that of *behavioural altruism*, which refers to the type of behaviour defined by its consequences that are not measured in fitness, have been presented in literature (Voorzanger 1994; Wilson 2002; Kokkonen 2003; Clavien & Chapuisat 2013). I will introduce the fourth notion of *agentive altruism* as the primary folk-psychological notion of altruism which should not be conflated with either psychological or behavioural altruism.[160]

---

[159] Some authors recommend separating altruism from prosocial behaviour because of the danger of conceptual confusions (for example, Hawley 2014 and Eisenberg & Spinrad 2014), but since it has been common to use "altruism" very broadly and it has been philosophers' practice to make distinctions between various notions of altruism instead, I will follow the "many altruisms" route.

[160] Christine Clavien and Michel Chapuisat (2013) distinguish between psychological, reproductive, behavioural, and preference altruism. Their definitions

6.1. Reasons and Causes to Help

Let us return to the example of Beatrice rescuing Amos from drowning once again. We can give this event several explanatory descriptions (only some of which are relevant here). First, we can give it an agentive description. If Beatrice saved Amos intentionally, instead of, say, falling into the water and dragging Amos with her accidently, we can identify the action as saving Amos. This is intentional helping, so we can identify this behaviour as helping as a *type of behaviour*. I will call a form of behaviour **behavioural altruism** if it is such that its results help another individual and this helping is not accidental. I will provide behavioural altruism with a more sophisticated definition later, but it is sufficient that the behaviour is an intentional act and the intention is to help. Beatrice's reason to save Amos may be instrumental only, even if the action is intentional helping – it could also be that she expects a reward of some sort or has some other *ultimate* goal that motivates this action, such as appearing heroic for the social benefits this brings. We would perhaps not want to call this a genuine act of altruism. I will stipulate that for the act to be genuinely altruistic in the *agentive sense*, or to be **genuine agentive altruism**, its ultimate goal (as perceived in a folk-psychological or action-theoretical framework), has to be the help that is provided (saving Amos, for instance). That is, the agent's *ultimate desire* in the situation is to help.[161] I will call

---

are different from mine, but why they distinguish behavioural and preference altruism from psychological and evolutionary (which they call reproductive) altruism is similar to my reasons for distinguish between behavioural and agentive altruism. The most important difference is that their preference altruism, which they define as an "action [that] results from preferences for improving others' interests and welfare at some cost to oneself" (Clavien & Chapuisat 2013: 131) is ambiguous with respect to what "preferences" refers to.

[161] Robert Richards (1987) uses the term "action altruism", which is applicable to both humans and animals and includes an idea of psychological motives. It is not the same as agentive altruism here and probably closest to what I will later call *quasi-intentional biological altruism*.

agentive pseudo-altruism an act that is intentional helping but only instrumental for something else that is self-serving (and intended as such), such as enhancing reputation, or in which the ultimate desire is something that is otherwise not directed to the receiver of help, such as a desire to hold on to a principle.[162]

Agentive altruism is conceptually separate from **psychological altruism**, which is defined by a *motivational mechanism* that causally guides the action, if we make the distinction between agentive and psychological as I have suggested. The altruism debate in social psychology has been about how to explain helping behaviour psychologically. In this context, "altruism" often refers to helping behaviour itself, regardless of how it is explained (in which case it is altruism in the behavioural sense), but sometimes it refers to an altruistic motivation in contrast to a selfish motivation to help. Some egoist theories explain helping behaviour (or at least much of it) with something that the agent achieves by helping (status, credit, social capital, gratitude, avoidance of direct or indirect sanctions etc.). Some others refer to internal self-serving states instead, such as pleasure from helping, avoidance of distress from seeing another person in need of help, or avoidance of guilt and shame from not helping. (See Batson 1991 & 2011 for a general review.) The first group of explanations make helping behaviour agentive pseudo-altruism (the ultimate desire to help

[162] There is a further issue regarding how to conceptualize common goals. The issue of joint action was briefly discussed at the end of the previous chapter, but it should be mentioned here that in all the types of altruism discussed here (agentive, psychological, and behavioural) it is possible to distinguish between selfish, altruistic and joint goals – Margaret Gilbert (1994), for example, makes the three-part distinction about psychological motives. The variety of possible psychological motivation mechanisms is, however, much wider than this, as we shall see. We could also make a distinction between being motivated by the good of the group *including* oneself, and good of the group *excluding* oneself. In evolutionary contexts, this distinction does not work: joint action can be mutualistic or reciprocal, and it is an open issue in both cases whether the behaviour of a participatory individual is selfish or altruistic.

lies in something that the agent gains). The second group of selfish motives may still be a basis for genuine agentive altruism: even if we classify the internal mechanisms as selfish, the action itself, under an agentive description, might have no other goal than to help. For example, it is a mother's distress that mediates the helping of her child, and if there are no other desires than to help the child (under an agentive description), there is no reason to call this a selfish act. Nothing else is intended. For it to be psychological altruism, something else is still needed.

6.1.1. Psychological Altruism

In contrast to the egoistic motivation theory, *the empathy-altruism hypothesis*, defended most importantly by C. Daniel Batson (1991 & 2011), states that there are genuinely altruistic motives and the motivation to help shows *empathetic concern* directly. Batson (2011: 11) defines empathetic concern as "other-oriented emotion elicited by and congruent with the perceived welfare of someone in need". It could also be called *compassion*.[163] For Batson, genuine altruism is motivation guided by empathy. He defines it as (Batson 2011: 20–21):

---

[163] Batson (2011) also distinguishes empathetic concern from other meanings of "empathy", which includes the following: (1) "Feeling" other people's emotions or thoughts (as in mindreading by simulation). (2) Matching the other person's posture or response (also known as "motor mimicry" or "physiological sympathy"). This plays a role in the Preston & de Waal (2002) model of empathy that I will describe later. (3) Coming to feel as the other. (4) Taking the other person's perspective in the situation. (5) Imagining how the other feels or thinks. (6) Imagining what oneself would feel in another person's situation. 7) Being distressed when seeing another person in distress. Elisa Aaltola (2014 & 2018), in turn, distinguishes between five kinds of empathy: projective/simulative (putting oneself in the other's position), cognitive (representing the other's states of mind), affective (reverberation of the other's emotions and other phenomenal contents), embodied (intersubjective sharing of bodily expression), and reflective. All these "empathies" may be relevant to altruism in one way or another.

1) a motivational state
2) with the ultimate goal
3) of increasing another's welfare.

The motivational state has a *goal* – it is not simply a drive, and Batson uses the concept of desire here. At the same time, it is a "force" – it is a causal factor that pushes the agent and disappears when the goal is reached. The goal being ultimate means that the motivational force does not cease through an alternative route. This is the difference from a selfish mechanism for helping such as distress, for which the agent does not feel the need to help if there is an alternative way to get rid of the distress (cf. Sober & Wilson 1998). The third criterion distinguishes between egoism and altruism.

A couple of clarifications will be helpful. First, Batson uses agentive concept of goal-directed desire here, although he refersto a causal psychological state. This is to be expected, as it is in psychology generally: the ambiguity of folk psychology shows its influence. But the conceptualization of the state as *agentive* desire does not do any work here. The psychological state of being motivated can be connected to an external target or goal without introducing a reason. What is required is that there are triggering conditions for both starting and stopping the motivational state (incentive salience), and this motivational state must relate to cognitive processing about the other person in need of help, which decides the need for help, how to do it, and when it is not needed anymore. This cognition can be a complicated representational system without the *drives* becoming any more complicated states. Even if folk-psychological conceptualizations are used to individuate the motivational states by representing their goals, including by the agent herself, it is not the agentive state of desire that matters, but the causally efficient motivational state on a psychological level. But for as long as we are talking about intentional action, the

Some of them might be necessary for perceiving the other person's distress or figuring out what would be good for them, for instance.

action needs to be intentionally helping – the altruistic consequences of behaviour cannot be accidental. This is not a part of the definition of *altruistic motivation*, but a part of what constitutes *helping behaviour* when we talk about humans.

Secondly, the issue is about the motivational mechanism that is a part of how the mind functions – not about individual acts as such. Empathetic concern is a habit of mind, a mechanism that is activated by perception of someone in need of help and produces an altruistic motivational state. It is not a property of an action or a behavioural trait. It is a *psychological disposition* that partly explains both helping behaviour and individual altruistic acts. An altruistic behavioural trait, in contrast, would be the *tendency to act* altruistically in the given context. If the empathy-altruism hypothesis is correct, it does not mean that humans act altruistically in every situation, and it does not even mean that everyone would act altruistically at least sometimes. Furthermore, it does not mean that the empathetic concern is behind all acts of altruism.

An example will illustrate both points. I am helping a colleague with her funding application. I enjoy taking up such intellectual challenges (unless it is my own application, in which case this joy is masked by a sizable stress factor), so I am happy to assist. I have also bonded with this colleague strongly enough that the friendship-related altruistic concern produces purely altruistic motivation. I may recognize both motivational forces, and I might use both as a basis for attributing myself reasons to help her with the application. But it may be that only one of these motivational states (pleasure or altruism) is behind my helping. If this is the case, removal of this factor would stop me wanting to help – this is a causal explanatory claim. Both might be effective motivating states, but only one of them strong enough for me to use my time and energy on my colleague's application instead of doing my own work. My intention is to help with the application (without further goals), in both cases, and both reasons are valid in rationalizing my action, as well as simply saying that I helped her because I wanted to. But only one motivation mechanism is active.

Note that the difference between selfish and altruistic motivations is qualitative, but they are not the only ways to motivate the action, and several motives may be active at the same time. For example, a mother helping a child may both have an empathetic concern and be distressed at the same time, as well as feeling social pressure to be a good mother, et cetera.

As mentioned above, there has been some controversy as to whether genuine altruistic motives exist. The proponents of altruistic motivation theory, such as Batson (1991 & 2011) and Sober and Wilson (1998), promote a pluralistic motivation theory that allows both selfish and altruistic motives. A pluralistic theory does not need to maintain that there are *any* situations where humans prefer others to themselves. Sober (1989) distinguishes between four motivation types: pure egoism, pure altruism, self-over-others, and others-over-self. If Beatrice is a pure altruist, she will always choose to help Amos, regardless of the consequences to herself. If she is a pure egoist, she does not pay any attention to his needs. If she has a pluralistic motivation structure, she might prefer either herself (self-over-others) or Amos (others-over-self). In the first case, she will always choose an action that benefits herself, but if there is a choice that makes no difference for herself but does for him, she will choose the one that is beneficial to him. In the last alternative, she will think of him first and only then herself. We can call the kinds of altruism that include self-sacrifice as an element **strong psychological altruism** and the kind in which the others matter but only after the self **weak psychological altruism**.[164] Even weak altruism is a kind of altruism, since the welfare of the others is a genuine goal (a factor triggering a motivational state), not a means to something self-serving – it is only that in case of contradictory

[164] Kitcher (1997: 285) also makes a distinction between strong and weak psychological altruism. By "weak altruism" he means only that someone else's wellbeing is considered, but the contrast (strong altruism) is not clear. Probably the distinction is approximately the same as the one made here.

goals, they prefer self-serving goals.[165] This distinction between different strengths of altruism can be applied both to psychological and behavioural altruism. In the latter case the quality of the motivation (whether it is empathetic concern or a selfish internal motivation mechanism of some sort) is left open, but the behavioural tendencies are the same.

Another clarification concerns increasing another person's welfare. The important part of this condition is that the incentivising factor is the good of the other person. It should not be a part of the definition that welfare is actually increased. The attempt to help may cause only harm, but this does not make the motivation to act non-altruistic. Furthermore, the measure of the good of the other could be various things, such as welfare as perceived by either party, or the recipient's pleasure, preferences, interests, achievement of their goals *et cetera*. An altruistic act may be directed to what the recipient of altruism wants, or it may be paternalistic. The main thing is that the aim is something that the agent thinks is good for the recipient. Although the paradigmatic example of helping behaviour and altruism is a situation in which the target of help is in distress of some sort, the point is about attempting to *increase* the well-being of the recipient. This does not need to involve something being wrong – giving a free ticket to a jazz concert just to be nice to the recipient is an altruistic act under any circumstances and a relief of distress of the recipient only in some. The actual results of the action do not matter – the helping intention (with the right kind of motivation mechanism behind it) does. The good intentions may lead to suffering (the jazz concert might turn out to be suffering for someone who does not enjoy complex music), but it is the intention that matters.

---

[165] The psychological debate is about the existence and the nature of altruistic motivation. If it does exist, however, whether the behaviour is guided by selfish, altruistic or a hybrid system is also contextual: the context, type of behaviour, the stakes, and the recipient, but also the individual and his or her personality matter. Some people are more helpful than others (on the behavioural level), and this may reflect their psychological tendencies.

With these modifications and clarifications, we can define **genuine psychological altruism**, for now, as applied to a particular act, as an intentional act that

PA 1) intentionally increases another person's welfare, pleasure, preferences, or something else that the agent believes to be positive for this person,

PA 2) does not require a further intentional goal for the action (whether such a goal exists as well), and

PA 3) is guided by an altruistic motivational state, which is made altruistic by being triggered by the perceived need or possibility of increase in the other person's welfare, pleasure, or something else positive to the person.

If the third condition is not satisfied, the act is **psychological pseudo-altruism**. It may still be *genuine agentive altruism* if the second condition is satisfied. If not, it is *agentive pseudo-altruism*. In the context of intentional action, psychological altruism is a form of agentive altruism that has a further criterion for the psychological motivational states that guide it. But things get more complicated when we address behavioural traits.

6.1.2. Behavioural Altruism in Psychology

For now, it is sufficient to characterize behavioural traits as tendencies to act in a certain way in a specified situation. As discussed previously, folk-psychological descriptions are one way to understand what instances of behaviour go together, but this may be misleading for some purposes. Social psychologists, or anthropologists, for example, might discover ways to classify types of helping behaviour on other epistemically justified grounds. For example, various relationships, such as family, friendship, colleagueship, tribal membership

(literally or figuratively) may be relevantly different contexts for otherwise similar behaviour, and the conditions for collaboration may be very different. Furthermore, although I have been calling all prosocial behaviour simply "helping," developmental psychological studies seem to show that there are finer differences. Sharing (which involves understanding the other's material needs) and comforting (which involves understanding the other's emotions) are psychologically different from helping with goals (which involves understanding actions), which might reflect differences in types of motivation or in the accompanying cognitive skills (Paulus 2018). Differences like this may be only partially acknowledged by the participants themselves. Some differences may also be constituted by cultural practices.

When we move to behavioural traits, we move from explaining actions to explaining robust behavioural tendencies. We are not explaining people's behaviour in the context through what they were thinking and what they intended to accomplish, but why they tend to think in certain ways and want certain things – or behave in some way even if they do not think at all. Even if all instances of a given behavioural trait were rationalizable actions, we would cut the direct connection with the agentive descriptions. There are folk-psychological ways of talking about behavioural tendencies, such as personality traits, and desire-like states and reasons can be standing dispositions, but agentive descriptions are about specific goals or reasons and usually about specific actions. If a person – let us say Beatrice, once again – has altruistic *tendencies* towards her friends, we might say something like "she always wants to help her friends," but her intentional acts are intentional acts to help a specific friend in a specific matter. We would not say, within the folk-psychological framework, that she is guided by her desire to *help her friends* in all these situations. Making such statement sounds more like a statement of a principle: Beatrice wants to be a helper of friends. These two things may be true at the same time, of course. But the latter would make the specific friend in need – Amos, again – in the given situation a placeholder for friends in the abstract. Beatrice's desire is to help friends, and her helping

Amos is derivative of this. Instead, what we mean if we say that Beatrice always wants to help her friends is that, in any given situation when Beatrice observes a friend in need of help, she has the tendency to desire to help that friend in that situation with no further goals.

Beatrice's tendency to help friends in need is her **behavioural altruism** towards her friends. Behavioural altruism is a characteristic of a trait, not of an action. It is a behavioural tendency to behave altruistically whether this is guided by a psychologically altruistic motivation mechanism or not. This is the helping behaviour in social psychological research, and this is often the reference of "altruism" in these contexts, instead of empathetic concern. We can define **genuine behavioural altruism** as a behavioural tendency to help without any further goal for the action. The psychological motivation mechanism for this may also be altruistic (empathic concern), but it may be a psychologically selfish mechanism (such as distress from perceiving someone else in distress, guilt, need for affiliation, avoidance of social sanctions, or enhancing one's own social status) as well. It could also be something completely different, such as following an internalized social norm as an action script, following a moral judgment in the situation, compliance with requests, or goal contagion. These are not conceptualizable as either altruism or egoism. (See Batson 1991 & 2011; Paulus 2018.) **Behavioural pseudo-altruism**, in contrast, would be the kind of helping behaviour that is conditional to some other external goal (that the act is instrumental to achieving) or a personal benefit, such as reputation.

Focus on behavioural traits raises the question of what makes a trait. As previously discussed, this depends on the question setting. What we are interested in here, ultimately, is the evolutionary approach, and evolutionary functionalism. The starting point is that it must be repeated, robust pattern of behaviour. As discussed earlier, human behaviour has invisible future goals and backgrounds, and it is loaded with cultural significance. It cannot be defined in a behaviourist way. Even *pancultural* behaviour (when the trait is given a functional definition) may be culture-specific in its phenotypes. For

example, if we give friendship a definition that refers to reciprocal altruism, the types of interaction and the scope of altruism within them may vary from culture to culture. Moreover, human behaviour is context-sensitive, and phenotypically similar patterns may be functionally different across cultures. This also means that cultural meanings, not only folk-psychological interpretations, need to be considered in the definition of any particular behavioural trait, even in the evolutionary functionalist framework. For now, the discussion on altruism in general, this complication does not arise. With all these caveats, we can use the intentionality of human behaviour as a guideline in choosing which parts of behaviour are part of some behavioural trait rather than accidental results, even if we do not categorize the traits based on the reasons given to these actions.[166] Defining a behavioural trait for an animal lacks this aspect, making it more straightforward in some senses and more complicated in others – I will turn to biological altruism now, after which I will return to how all these different kinds of altruism are related, and why all this matters.

6.2. Biological Altruism

In biology, there are three notions of altruism in play. First, there is **evolutionary altruism**, which is defined by the fitness consequences of the trait (which needs not be a behavioural trait, but this is what we are interested in here). Second, there is a concept of **behavioural altruism** – for example, Frans de Waal's (1996 & 2009) concept of altruism in his discussion on prosocial behaviour in animals is that of behavioural altruism. In addition, de Waal and his collaborator Stephanie Preston present an empathy model that matches Batson's empathetic concern model in psychology and is therefore a theory of **psychological altruism** when used to explain altruistic behaviour

[166] This is not always the case. Some behavioural traits may include unintentional doings. We do not need to worry about this yet.

(Preston & de Waal 2002; Preston 2007; de Waal 2009; Batson 2011). I will discuss all these altruisms in the biological context, but I will start with a couple of remarks on the concept of fitness, which is central to the evolutionary concept of altruism.

6.2.1. What Is Fitness?

Fitness is a measure of an organism's capacity to produce offspring.[167] It is not a measure of the actual number of offspring – identical twins have the same fitness even if they have a different number of offspring.[168] Fitness depends on the phenotype: both on the individual's capacity to produce offspring as such (fertility and ability to attract mating partners) and adaptivity to the environment. It would be wrong to say, however, that individual phenotypic traits *cause* fitness – fitness supervenes the phenotypic properties. It is an abstraction from the causal basis that defines how many offspring the organism is likely to produce. Fitness does not depend only on the traits, but

[167] There are two different ways to think about the number of offspring: short-term or long-term (Sober 2001). Short-term fitness is the number of offspring that continue reproducing in the next generation, but long-term fitness refers to the number of offspring in the distant future, or some other measure, such as how many generations it takes for the genotype to become extinct. The notion of short-term fitness is directly related to the mechanism of natural selection. Long-term fitness depends on short-term fitness and the ecological and population-level conditions. For example, male lions who replace the previous male of the pride kill the offspring of the previous male. This behaviour increases short-term fitness, but under some conditions (such as small population size and frequent overturnings), this will lead lions with this tendency into extinction more quickly. Long-term fitness id of use in ecology, for example, but as we are interested in the mechanism of natural selection here, the short-term notion is more adequate.

[168] Defining fitness as an actual number of offspring would also risk making natural selection explanations circular, but this can be solved with the propensity interpretation of fitness. See Brandon 1978; Mills & Beatty 1979.

also on how the traits work together in the configuration of the concrete individual that reproduces. The individual organism has only one fitness that binds the traits together (they have a "common faith"). The fitness differences between individuals are quantitative, but the phenotypic differences are often qualitative. The phenotypic differences can be quantified using the fitness difference between the *individuals* who share the different variants of the trait on the population level as a measure, but this does not make the differences between the *traits* quantitative. Furthermore, fitness also depends on the environment – the adaptivity of traits differs with the environmental variation, and the order of organisms measured in fitness may vary depending on the environmental factors, including other individuals of the population that the organism interacts with, and how the organism modifies the environment itself. (See Futuyma 1998, 366–371; Rosenberg 1978; Dawkins 1982; Sober 1984; Odling-Smee *et al* 2003.)

However, natural selection operates through the heredity of traits and their various contributions to the fitness differences between individuals. Fitness differences between variants are usually discussed and modelled as two (or more) phenotypes that differ only in this one trait, and the genotypes attached to the phenotypes are equally discrete types that are internally identical. This is a useful abstraction for modelling, but it is unrealistic in several significant ways. First, as stated above, the trait's contribution to fitness depends on the overall phenotype. Therefore, for sexually reproducing organisms, the fitness of a genotype is defined as the additive sum of the fitness of its alleles, which in turn is the average fitness contribution of the alleles in the entire population (Futuyma 1998: 368). Second, since the phenotype is a co-product of the genotype and the environment, identical genotypes may have different phenotypes. Having phenotypic differences between the same genotypes in different environments may even be adaptive and a product of selection for functional, stimulus-sensitive plasticity (see West-Eberhard 2003). Yet it is the genes that are replicated, not the traits. These considerations are one reason to think that it is the genes, or the replicators, that are central to evolution,

and we should be interested in their average fitness in the population – to take a gene's eye view to the evolutionary process. However, this is a problematic solution. The concrete genes that get copied and participate in the developmental processes, the *D-genes* (Moss 2004), are DNA sequences.[169] The "genes" that are referred to in population genetics, the *P-genes* (Moss 2004), are abstractions of DNA-genetic level factors on the population level that cannot be reduced to individual D-genes. They are not what get copied. This entails some problems.

First, the developmental processes are complicated interactions between the existing characteristics or developmental stages of the developing organism, the genetic factors, and the environmental factors, and the whole process takes place in stages (see Oyama, Griffiths & Gray 2001; Wolpert *et al* 2010; Griffiths & Stotz 2013). Second, the environmental factors may be alternative routes to reproduce traits and the evolutionarily significant differences (see Oyama, Griffiths & Gray 2001; Jablonka & Lamb 2005; Pigliucci & Müller 2010). Since P-genes are abstractions from this process, it makes no sense to consider them the objects of evolution. That would be a reification fallacy. They are simply a shorthand for talking about the phenotypic differences that the genetic differences are responsible for on the population level – which is the relevant explanatory perspective on the population level but does not make P-genes entities. On the other hand, if we consider D-genes, the actual things that get replicated, to be the replicators of the replicator-based interpretation of evolution, we do not have any reason to differentiate between them and the other causal factors on which the process depends. After all, the process of natural selection requires fitness-relevant differences in traits between the individuals, and that they are reproduced – in other words, that there are relevant phenotypic hereditary patterns – but not that the process of reproduction

[169] Furthermore, only some of these genes "code" for a trait in the way proposed in the sequential hypothesis, from DNA through RNA and protein production to phenotypic traits. There are regulatory genes, for example, that only influence how other genes are read. (Wolpert *et al* 2010; Griffiths & Stotz 2013.)

is made of replications like this (see Godfrey-Smith 2009).[170] Other ways to conceptualize the evolutionary process than the replicator-based ontology have been proposed. For example, the *Developmental Systems Theory* views reproduction as the *reconstruction* of similar phenotypes from the developmental resources, genetic and otherwise, some of which are inherited and some of which are not (see Griffiths & Gray 1994; Oyama, Griffiths & Gray 2001). In Peter Godfrey-Smith's *neoclassical individualism*, the level of analysis is an organism that produces a similar organism with high fidelity (Godfrey-Smith 2009). They both have a more holistic view of individuals and a more modest role for genes.

I will try to stay neutral on the ontology of evolution, but the reasons for proposing alternatives are valid (for example, genes are not the only form of replication that should matter) and I will take them into account in my discussion later. One complication is rethinking the idea of *inclusive fitness*. I will return to the recent debates on the relationship between kin selection and group selection, but for now, it is sufficient to say that I have been referring only to direct fitness in my discussion so far. Replacing direct fitness with inclusive fitness in defining altruism would seriously disturb the framing of the very problem of altruism in evolution anyway (see Sober & Wilson 1998; Okasha 2006; Birch 2017).

The general lesson is that talk about fitness of genes should be dropped, and fitness-talk should be limited to the interactor entities (individuals, superorganisms, and possibly groups of other kinds). This affects how we should approach kin selection and reciprocal altruism – solutions to the problem of altruism that use the gene's point of view. It might be tempting to replace genes with the phenotypic traits they are associated with as the sub-individual units of fitness, and to think about the fitness of an individual as the sum of the fitness of its traits. In fact, it is common to talk about the fitness of traits in the

[170] The reproduction must involve, however, some units with a "minimal size" and fidelity in replication, or there could be no selection (Depew & Weber 1996; Godfrey-Smith 2009).

philosophical literature. For example, Elliot Sober does so and defines the fitness of a trait as the average fitness of the individuals who have the trait (see Sober 1984 & 2001). However, since fitness is the measure of something that reproduces, it would be a category mistake to apply it to *part* of a phenotype. The perspective-dependence of what counts as a trait in the first place and the overlap of traits also make this much trickier than if we were talking about discrete alleles – especially in the case of behaviour. Furthermore, because of pleiotropy and other developmental connections between traits, as well as structural connections and coincidences, the fitness of a trait as a statistical measure only would make no distinction between selection *of* and selection *for* (Sober 1984), which is a crucial part of understanding natural selection explanations. Therefore, the expression "fitness of a trait" should be understood as a shorthand for "the trait's contribution to the fitness of the individual that has it," which is a completely different thing.

The final remark on fitness to be made here is about the distinction between *absolute* and *relative* fitness. Absolute fitness is a measure of reproductive capacity, relative fitness a measure of the reproductive capacity relative to the variation of reproductive capacity in the population, with a value between 0 and 1. Relative fitness is simpler in modelling and fitness compared to other individuals' fitness is what matters for what gets selected. There are, however, differences between the definitions of evolutionary altruism, depending on which notion is used. Evolutionary altruism is defined as a property of a trait that is suboptimal for the individual (compared to the other variants) but increases some other individual's fitness. If we use relative fitness in the definition, *all* suboptimal behaviour will be altruistic, since they increase the relative fitness of individuals who have more optimal variants. One suggestion for a solution to this could be to restrict the traits to only those that do something *directly* for the benefitting individuals, but this would rule out all structural properties (such as the hooked sting of honeybees) that are (and should be) included in the general problem of altruism. Furthermore, cooperation tendencies that *increase* fitness, but less than an alternative trait would, would still count

as altruism. Another suggestion could be to include a reference to group level in the definition. This is done in some treatments of altruism under multilevel selection (MLS) models (for example, Peressini 1993; D. Wilson 2002) and it would fit the evolutionary functionalism model assumed here for the individuation of traits. There are two problems here, however. First, this can be done only within the MLS framework, which is contested, and the phenomenon itself should be framed neutrally. Second, only those apparently altruistic traits that can be *explained* by group selection and not by individualist models would be included in altruism, which would be unintuitive and make the explanations of altruism group selection explanations by definition. This means that we should define altruism in terms of absolute fitness (as done explicitly by Rosenberg 1992, Kitcher 1997, Sober & Wilson 1998, and Nunney 2000, for example) even if relative fitness values are used in the calculations.

6.2.2. Evolutionary Altruism

It turns out that the choice of the concept of fitness has consequences. Altruism defined in terms of relative fitness can evolve in contexts where altruism defined in terms of absolute fitness cannot (Mitteldorf & Wilson 2000; Kerr *et al* 2004). David Sloan Wilson refers to the difference by distinguishing between *strong* and *weak* evolutionary altruism (Mitteldorf & Wilson 2000; Sober & Wilson 2000b). But what Wilson shows is only that suboptimal traits can evolve if their existence benefits the other individuals in the group. This is an interesting result of using an MLS-model, but it does not address altruism specifically. There is, however, another possible distinction between a weak and a strong version of altruism: the difference between a trait that increases another individual's fitness and decreases the individual's own fitness and a trait that increases the other's fitness without decreasing one's own. If we define this in terms of relative fitness, the trait must increase the individual's own fitness at least as much as it increases other individuals' average fitness. If we use absolute fitness as a

measure, as I think we should, weak altruism still decreases the relative fitness of the individual. This highlights the "self-sacrificing" nature of helping from the evolutionary point of view even when there is nothing self-sacrificing as such in the behaviour itself.

Benjamin Kerr, Peter Godfrey-Smith, and Marcus Feldman (2004) have shown that the various definitions of altruism that exist in literature behave differently in modelling evolution. The definitions they discuss vary in who is considered the benefitting party (other individuals, the group complement of the altruist as a block, or the whole group including the altruist) and in whom the costs of altruism are compared (the individual fitness across the population or the local group only). Different conditions must be satisfied for "different altruisms" to evolve. This means that they are descriptions of slightly different evolutionary properties of the traits. Indeed, as Kerr, Godfrey-Smith and Feldman (2004) show, none of the definitions captures a sub-category within the extension of another definition, but they are genuinely different concepts. This is only partly a definitional issue, however. I will discuss this using a definition of generalized evolutionary altruism presented by Anthony Peressini (1993: 584–585) as a point of departure:

1) The altruistic trait is less fit than *at least one* of the alternative traits, when *all* traits are present.

2) The relative frequency of an altruistic trait is *more than zero* in an *optimal group.*

3) The altruistic trait must be causally responsible for the transfer of fitness, which explains why the first two conditions are satisfied.

Peressini's first condition is needed to generalize the applicability of the concept to all variants in the population. Altruism is usually defined and modelled in a system of two variants, altruistic and selfish. This hides the relativity of altruism and selfishness as evolutionary concepts: whether a trait is altruistic or selfish depends on other

variants. If an altruistic trait is selected because of group selection, for example, and the selfish counterpart disappears, this gets hidden. This is why we need a comparison to all alternatives.[171] The trait can be more beneficial to the individual than *some* alternatives, as long as there are alternatives that are more beneficial. It cannot be a requirement for altruism that it is the worse alternative, and this is why the trait must be compared to all alternatives. The second condition expresses the benefit for other individuals. *Optimal group* has the dispersion of alternative traits that increases the fitness of the whole group more than any other dispersion. If the first condition is satisfied, the second condition can be satisfied only by contribution to the fitness of others. The third condition is to rule out coincidental correlations.

The second condition is problematic because of its reference to groups, for the reasons mentioned above. Furthermore, it follows from the condition that only those traits that maximize the fitness of others are altruistic: a trait that increases the fitness of others, but less efficiently than an alternative trait, would not be a part of an optimal group and therefore would not be altruistic. In other words, evolutionary altruism as defined by group benefits does not define altruism but another property that we might call *evolutionary groupism*: being such that it maximizes the fitness of the group (despite being suboptimal for the individual). It may turn out that being evolutionarily groupist is the only way to be evolutionarily altruist, but this is an

[171] "All alternatives" is an ambiguous expression. The range of alternatives depends on the explanatory context the same way that all evolutionary comparisons between traits do, and it has context-dependent constraints. The relevant alternatives to compare are usually those that actually exist, have existed, or whose appearance would be reasonable to expect. The last option may have stricter or looser and more or less realistic constraints, depending on the theoretical interests. A concrete consequence is that a trait could be altruistic with a wider comparison of alternatives but selfish within the historical variation. This is a possible explanation for some traits, but we can consider this an exception while discussing the mechanisms that *actively* produce altruistic traits.

explanatory relationship. As a definition, it fails. The second condition needs to be replaced by a condition that refers to fitness effects in *other individuals* instead. Furthermore, the trait does not need to benefit the whole group (whether defined as a group or as individuals) to be altruistic, if the core meaning lies in increasing fitness in others.

In the light of these considerations, I define **evolutionary altruism** as follows: a trait is evolutionarily altruistic *iff*

EA 1) at least one of the alternative traits would make the individual fitter in the *absolute* sense of fitness when all traits are present, and

EA 2) the trait contributes to the fitness increase of at least one other individual.

The third condition is unnecessary since causality is included in both parts of the definition. It should be noted that this definition is only meant to pick those traits that are altruistic. In modelling the evolution of altruism, for example, relative fitness should still be used, and it is probably less complicated to model altruism as a groupist trait instead of "only" altruistic, as in the MLS paradigm. The differences in the definitions are a problem for theoretical completeness, but also highlight that there is probably more than one explanation for evolutionary altruism if it exists.

This definition, like all definitions of evolutionary altruism, lacks a definition for a trait. Consider the following example. There is an environment with only one species of predators that hunts a certain prey animal. For example, the predator species is the only species large enough to hunt prey animals above a certain size on a regular basis. Now, the members of the predator species also have some trouble with the large prey animals and only succeed in injuring them seriously without capturing them on a regular basis. This is a failure in what they are trying to do. However, the injuries of the prey make it significantly easier to capture later. This benefits all the members of the species at the same region. This trait – injuring without capturing

– seems to be an altruistic trait according to any definition of evolutionary altruism, at least if we stipulate the fitness effects in this thought experiment in the right way.

But intuitively this is not an example of an animal behaving altruistically towards its conspecifics. The animal is failing at something that it is attempting to do and would do if there were no external constraints. This example is a variation of one provided by Jack Wilson (2002), who criticizes the definitions of evolutionary altruism of making some behaviour "accidentally altruistic" (see also West *et al* 2007b). He also points out that concentrating only on evolutionary altruism ignores apparently altruistic behaviour that can be explained as evolutionarily selfish and calls for something analogical to intention that could be used to define what the behaviour is about (see also Voorzanger 1994). It is therefore necessary to clarify the definition of behaviour in biology, and to have a behavioural-level trait definition for altruism.

6.2.3. Behavioural Altruism in Biology

According to the evolutionary functionalist approach for individuating traits that I proposed in the previous chapter, a behavioural trait is the collection of behaviours that jointly achieve something that has positive fitness consequences in the environment. We can define it as follows:

BT1) There are repeatable patterns of behaviour that have a robust correlation with an achievement (such as a change in the environment) and a mechanism explaining the connection between the behaviour and the achievement.

BT2) The achievement has a positive correlation with the fitness of an individual performing the behaviour.

BT3) This behaviour is repeated as a behavioural tendency.

The first part of the definition binds the behaviour with an achievement. The second part distinguishes achievement as something that the behaviour is about (using its adaptive value) rather than the mere consequences of behaviour. There may be several ways to bring about the achievement. The various behavioural patterns may be classified as separate traits or the same trait, depending on the explanatory interests and theoretical contexts. The same instance of behaviour can be a part of several behavioural traits. The mechanism that explains the achievement may consist of the bodily behaviours only (if the achievement is a direct result of the bodily movement) or include environmental factors, including responses from other individuals. It may be a combination of various things that the individual does and various capacities that it possesses, such as in the example of honey buzzard. Everything that is a part of the individual's behaviour and is necessary for the mechanism that produces the achievement is a part of the behavioural trait. The third part of the definition distinguishes a behavioural trait from a single instance of behaviour. The difference between a few instances and a fixed tendency is gradual, which means that it cannot be quantified in the definition. What matters is that the behaviour is common enough to have the positive fitness benefits.

This account does not state that the trait must be optimal, only that there is an evolutionary "purpose" for why an organism does what it does: a positive fitness effect through what is achieved. If a behavioural pattern does not have a fitness effect that unifies it in this way, it does not have an evolutionary function and it is not a trait that we could understand from this perspective. This requirement rules out characteristics that are harmful, but it does not rule out altruism. First, the evolutionary function of a trait may involve long-term consequences that are beneficial and depend on other factors, including other individuals. The point of kin selection and reciprocity as explanations lies in the indirect fitness benefits for the organism of its "apparently altruistic" (or behaviourally altruistic) behaviour. As I will argue shortly, this implies that we should allow the behavioural traits under evolutionary explanation to be non-individualist. These traits

are usually thought to be evolutionarily selfish although behaviourally altruistic, but even if they are evolutionarily altruistic, they expand the proximate-dimension analysis to multiple individuals and the fitness effects to all of them. Furthermore, if group selection is a theoretical possibility, we can approach the evolutionary functionality of a trait from this perspective, which allows evolutionary altruism to be evolutionarily functional. This may sound odd after I have just criticized the use of group selection in the analysis of evolutionary altruism. Some clarification is in order.

I have been explicitly pluralistic about how to define a trait. We can define a trait from some other perspective than evolutionary functionalist and discover that it is evolutionarily altruistic. In principle, we can define any robust pattern of behaviour as a trait. The problem with this regarding evolutionary explanations is that they can be completely random: there is no point in wondering how an altruistic form of behaviour in some context evolved, if it turns out to be a contextual consequence of behavioural tendencies that are beneficial to the individual choosing that particular strategy in most cases, for example (see Johnson, Stopka & Bell 2002). Psychological structures that are overall adaptive or developmental pathways that connect different types of behaviour can make apparently altruistic behaviour be only a side-effect. But as I have argued, we also need to have a way to break the behaviour down into tasks (with achievements), since these are what matter to why the capacities and tendencies behind the behaviour get selected in the first place. Therefore, we need a definition of behaviour that is connected to an adaptive function.

If we use some other criterion for defining what a trait is, we may conclude that the trait does not have an evolutionary function at all. Not all traits do, of course. For example, there is no evolutionary reason why giant pandas cannot digest most of the food they eat or why they are so reluctant to have sex. These characteristics may have an evolutionary explanation, albeit an evolutionary historical explanation that is not adaptationist. We could always take an evolutionary historical perspective and contrast adaptive and non-adaptive histories. This is

not, however, the route that I take here. First, the subject matter of the discussion here is evolutionary functionalism specifically. Second, my point about the usefulness of (ahistorical) evolutionary functionalism in understanding behaviour is precisely that it is a way to understand what something is about. This is not an empirically adaptationist stance. It is about structuring what the organism's design is and what it does. Let us think about the giant pandas a little bit more. Eating as such has an evolutionary function, and we can break the design of the panda's digestive system and eating behaviour into functional parts of the holistic system that has this function. Some details of this system are "bad design," which should be replaced by something more functional. But we can only say that something is badly designed in the context of approaching the system as one that has a design from an evolutionary perspective (see Orzack & Sober 1994a; Dennett 1995). The panda's digestive system is almost a reverse situation with whales having lungs: the design of the whale's respiratory system is very well adapted for being a sea creature with lungs, except for the premise of having a sea creature with lungs in the first place, which is not a very bright idea in terms of design. In the case of the giant panda, the behaviour is adapted to the environment (there is plenty of bamboo) but the digestive system is not adapted to that particular food very well. Still, the trait of interest here is the digestion of bamboo, which we identify with eating bamboo having an evolutionary function in the panda's design. It just happens to be badly instantiated.

In this example, bamboo digestion is a trait, and the alternative physiologies are *mutually exclusive variants* to achieve the same thing. Some possible physiologies do not achieve this very well, but the "achievement" that gives the function of the trait is fixed by the fact that the pandas are eating bamboo, which has the function of obtaining nutrition. In the case of behaviour, different behavioural strategies are not always "aimed" at achieving the same thing, in the sense defined above, but what makes them alternatives is that there is something of fitness value in the situation. There are different mutually exclusive behavioural choices with mutually exclusive achievements

regarding the same behavioural context, which determines the fitness gains and losses. For example, the comparison between selfishness and altruism is between alternative behavioural variants in the same social context. They are comparable only in relation to something that is achievable. Furthermore, we cannot restrict the comparison to behaviour in one context only if we wish to talk about behavioural traits, not just behaviour. We need to think about the more general function of the behaviour in the specific context from the point of view of the organism's form of life. In doing so, we can cluster all the *mutually exclusive variants* as a *contrast class*, including doing nothing. This class can be considered a trait in a broader sense – behaviour (whatever variant it takes) in relation to something at stake that is valuable in fitness in the context. This is analogous to digesting bamboo, and the behavioural traits in the narrow sense (what is done in this context, the manifest variant) are analogical to the alternative physiologies. In other words, we can use the evolutionary functionalist approach to define an actual trait but also to define a theoretical contrast class, which is essential for knowing the adaptive value of the variant in contrast to its alternatives.

Our interest in modelling the competition lies in discovering the population-level dispersion of the alternatives: which of the alternatives becomes dominant, or which frequency becomes stable – if there is an equilibrium in the first place (see McElreath & Boyd 2007). In modelling social evolution, the relevant alternatives are compared against each other. In the narrow sense of a trait, different traits are being compared, but in the broad sense of a trait, the alternatives are variants of the same trait. Identifying the evolutionary significance (the possible achievements) and figuring out the optimal variant are separate things to evaluate about the trait. We can identify the evolutionary function of digesting bamboo and discover that pandas are very badly adapted to it. We can also identify a social behavioural trait according to what the evolutionary role of that kind of social interaction is and discover that the actual behaviour is not optimal at all. In addition, it can increase other individuals' fitness instead. If this is

what the behaviour is systematically correlated with, it is its "achievement". It is not defined through its own positive fitness effect, but through it being an alternative to a trait that has an achievement with a positive fitness effect. An altruistic trait is a trait in an evolutionary functionalist framework because it is a variant to a selfish trait that defines what is at stake in the behavioural context. That said, the behaviour that appears to be altruistic may turn out to be beneficial in the end and, therefore, have an evolutionary function. This is the case with apparently altruistic behaviour that can be explained with kin selection or reciprocity-based models – and possibly with MLS models. I will return to this later.

Whether or not a trait is genuinely evolutionarily altruistic, it can appear to be, and this can be called **behavioural altruism**. Just like in a psychological context, it is a characteristic of a behavioural trait to be directed to someone else's wellbeing than the individual themselves in its goals. In psychological and biological contexts, the means to decide the directedness are defined differently (intentional goals versus the achievement) and the wellbeing is measured differently, fitness being the important factor in biology. There is no point in defining altruism as a phenotype with actual fitness consequences, however, for the reasons discussed above. The difference between behavioural and evolutionary notions of altruism is that one refers to the appearance of behaviour and the other to its actual consequences. From the evolutionary (functionalist) point of view, the appearances that matter are those that are connected to fitness. All this being the case, a behaviourally altruistic trait is a trait that *systematically transfers resources that are normally associated with fitness* with this species. For example, food has fitness consequences, so sharing food is behaviourally altruistic.

Behaviour must be non-accidental in order for it to be a trait and for the attribute of behavioural altruism to apply. However, there can be many different evolutionary rationales behind such behaviour, some of which we might not consider "genuine" altruism. It can be, for example, coerced behaviour, or done to appease an individual who

is higher in the social hierarchy, or it may be about trading goods.[172] If there is a selected mechanism for the altruistic behaviour and there are no further goals to be achieved in proximate terms, I call it **quasi-intentional biological altruism**. This is similar to the *agentive altruism* defined earlier, with the difference being that in agentive altruism, the ultimateness of the goal is defined by the agentive description (no further goals needed for there to be a reason to act) whereas here the functional design works as a quasi-intention, the "intention of the design."[173] In principle, this concept can be applied to both a particular behavioural instance and a generalized behavioural trait, but the interest in biological approaches to behaviour is almost always the latter. Biological altruism can be either evolutionarily selfish or altruistic, depending on its actual net fitness consequences. The same applies to behavioural altruism that is not quasi-intentional but induced. That is, the behavioural altruism (in general) can turn out to be either **evolutionary pseudo-altruism** (that is, it turns out to be evolutionary selfish) or **genuine evolutionary altruism**.

[172] As discussed earlier, Ronald Noë and Peter Hammerstein (1994, 1995 & 2016) have proposed that some animal social behaviour could be understood as trading of goods and services on "biological markets," especially in cooperation and partner choice. Whether this is an adequate explanatory device, much of the social activities of animals involve trading-like direct reciprocity (on the level of goods). This should not be confused with reciprocal altruism, which is "trading" on the level of fitness.

[173] It should be noted, once again, that the evolutionary "reasons" cannot be treated as analogical to agentive reasons. They are different dimensions of projecting teleology onto a complex functional system. Human behaviour, for example, might have both evolutionary and agentive reasons, and they would not compete or be directly related in any other way either. This applies to the notions of altruism as well, as we are about to see.

6.3. Kinds of Altruism and Why We Should Care about Them

Now it is time for a systematic presentation of the kinds of altruism discussed above, how they are related, and why distinguishing between them matters for explanations. The "basic" concept is **behavioural altruism**. I have used the same term in both psychological and biological contexts because I think the base concept is the same and its application requires various specifications in any case. Roughly speaking, it refers to behaviour that *increases a good of some sort in another individual*. I will provide a full definition later. We can apply this to both an individual occurrence of behaviour and a behavioural trait. The good that is increased needs a content, something to use as a measure. Using Beatrice and Amos as our protagonists again, the altruism of Beatrice's actions can be measured by the increase in Amos's *welfare*, *happiness*, or *accomplishment of his goals or preferences*. These three measures do not necessarily capture the same behaviour. We could also measure altruism in fitness, but this is a measure in a different dimension. These two dimensions lead into psychological and biological (or evolutionary) notions of altruism and further distinctions within these notions. However, these dimensions are not entirely separate.

6.3.1. Behavioural Altruism Elaborated

Which goods Beatrice seeks to maximize makes a difference: welfare, happiness, or preferences. There is another difference between Beatrice maximizing the good of a given kind the way *she* perceives it, the way she *thinks Amos perceives it*, and as an objective matter. There is also a difference between the helping being both intentional and altruistic and it being *intentionally altruistic*. In the latter case, it is *agentively altruistic*. The measure is bound to Beatrice's cognitive limitations: it can be altruistic only in the ways that Beatrice can think are altruistic. If the requirement is only that the instances of behaviour are intentional, and the behavioural trait overall is altruistic, the measure

can be objective, too. For example, if a parent has a tendency to scold their child for behaving badly, it may be a reaction born of annoyance and frustration but have an overall positive effect on how the child grows up.[174] This is not necessarily an important distinction in practice but has to do with how the concept of human behavioural altruism is related to the concept of biological behavioural altruism.

In animal contexts, but also in the human context from an evolutionary point of view, fitness is relevant. We can define another measure of behavioural altruism as *whatever typically increases welfare*, as discussed in the previous chapter. This is the core biological notion of behavioural altruism and one that defines the biological *phenomena* that are the *explananda* in the evolutionary explanations of altruistic behaviour. We can assume that some behaviours measured in happiness or by the accomplishment of agentive goals in human contexts are biologically altruistic in this sense, and welfare even more so, but this is not always the case. I will conflate the behavioural altruisms (the various ways in which it can be defined in human contexts and the fitness-related biological definition) on the level of analysis of the behaviour for the following reasons. Although the agent is limited in her ability to know what is good for the recipient, attempts to help nevertheless result in fitness increase more often than in decrease. The pleasure of the recipient is not necessarily fitness-increasing either,[175] but I assume that pleasure and distress are good proxies for the things that have fitness consequences and selected to be such. The same holds for helping with goals or preferences – not all goals and preferences increase fitness when accomplished,[176] but generally speaking

[174] This is not to claim that this is the case – scolding might have negative overall effects just as well.

[175] For example, the pleasure from eating sweet and fatty foods, which are selected preferences in scarcity, but are harmful in the contemporary Western abundance of sugar and fat.

[176] We may even have selected preferences that would decrease fitness if achieved, since they were selected for what we *actually achieve* while *trying* to achieve the goals. (See Sterelny 2003.)

helping to achieve them or to work towards their direction is fitness-increasing rather than fitness-decreasing for the recipient. Human behavioural altruism, therefore, *tends* to be biological behavioural altruism too. Selection for biological behavioural altruism in human contexts may mostly be selection of human behavioural altruism, and the evolutionary understanding of human behavioural altruism is the evolutionary understanding of biological behavioural altruism. But although I assume that the overlap between human and biological behavioural altruism is substantial enough to discuss them as generally the same, behavioural altruism is clearly distinct from both psychological and evolutionary altruism.[177]

As discussed above, it may turn out that the altruistic consequences of seemingly altruistic behaviour are not the (evolutionary or intended) function of the behaviour, but a side effect or an "error." We need a further criterion for connecting the behaviour with its altruistic consequences, and I have presented two ways to do it: through intentionality of helping, and through a selected mechanism to behave in a certain way, which I have called **quasi-intentional biological altruism**. In both cases, we can distinguish between behaviour that has the altruistic consequences as its *ultimate* goal and behaviour that is instrumental to achieving something else. We can call these four kinds of altruisms **genuine agentive altruism** (as defined earlier), **agentive**

[177] The deliberate conflation I am performing here is not the one that Neven Sesardic (1995 & 1999) considers to be the evolutionary problem of psychological altruism: that psychological altruism tends to produce evolutionary altruistic behaviour. Once the concept of behavioural altruism is introduced, psychological and evolutionary altruism become two different additional properties of behavioural altruism. It could even be the case that the strongest evolutionary cases for psychological altruism as a mechanism for producing the selected behaviour are those in which the behaviour can be explained as evolutionarily selfish (with kin selection or as reciprocal altruism) and the strongest cases for evolutionary altruism do not involve psychological altruism (the case being "altruistic punishment"). I have argued for this coincidental "paradox" elsewhere (Kokkonen 2003).

pseudo-altruism, **genuine quasi-intentional biological altruism**, and **quasi-intentional biological pseudo-altruism**. Again, these are not the same distinctions intensionally, and probably not extensionally identical either, but I will collapse them together in human contexts for simplicity. Agentive altruism is an agentive-level description. Quasi-intentional biological altruism is analogical to this in the sense that it is an interpretation of what an animal "attempts" to do. The direct equation of the two would be a case of the Freudo-Darwinian fallacy. We cannot do this in the case of an individual action. If we switch our focus to behavioural traits, however, we can think of both as behavioural altruism. In the human context, a behaviourally altruistic behavioural trait is a tendency to intentionally help under specific conditions. In the biological context, a behaviourally altruistic behaviour trait is, likewise, a tendency to be internally directed to help under specific conditions. In human contexts these may almost be equated, with three caveats. First, this can be done, even in principle, only in the cases where we are providing an evolutionary explanation for a behavioural trait. This is, however, what we are interested in here. Second, human behavioural altruism is connected to two kinds of intentionality, as mentioned above: the intentionality of the action that we consider altruistic, and the intentionality of altruism (whether this is the instrumental or ultimate goal for the agent – we are not discussing psychological altruism yet). This binds the nature of helping to the agent's understanding of the situation, including their epistemic constraints, which are not constraints on assessing something as behaviourally altruistic.

But, again, we are discussing behavioural tendencies here. The psychological mechanisms involved in these tendencies (including rational reflection and deliberation) are the basis for wanting to behave in certain ways. We can assume that, in most cases, people understand that they are helping someone when they intentionally act in a way that is beneficial to that someone. We can also assume that the criteria for us to evaluate an act as altruistic on the agentive level lie in the recognition of some psychological feature that we would also identify

as a part of the altruistic behavioural tendency as a proximate mechanism in the biological description. This does not need to be an altruistic motivational mechanism, though. Furthermore, all other psychological mechanisms (for example, completely instinctual ones) would also be fallible. So, once more, I assume that behavioural altruism, as defined in the section on human altruism, and behavioural altruism, as defined in terms of quasi-intentionality, overlap enough that it makes sense to use "behavioural altruism" as a unifying concept in discussing the evolutionary basis of altruistic behaviour. I do not propose that they are the same thing or that human behavioural altruism (consisting of intentional actions) is a type of quasi-intentional biological altruism. At minimum, quasi-intentional biological altruism involves actual intentional helping and therefore human behavioural altruism in human contexts is enough for this to be the most typical case of quasi-intentional biological altruism in humans. Therefore, without presuming a conceptual equivalence or full extensional co-occurrence, I will talk about behavioural altruism in this limited context (evolutionary perspective on human behaviour) as if it is both human behavioural altruism and biological behavioural altruism.

Now, we can redefine **genuine behavioural altruism** as a characteristic of a behavioural trait in which

BA 1) the person contributes to another individual's welfare or achievement of their preferences

BA 2) without this behaviour having the aim of increasing the person's own welfare or independent preferences.

"Independent preferences" means that the agent would not have these preferences without the recipient of altruism. The agent may prefer to make the recipient happy, for example, and the helping is instrumental to this, but the preferences would not exist independently. Helping may have positive consequences for the agent, for example, reputation or future reciprocity, but these cannot be what the behaviour is aimed at. For example, friendship as a form of

relationship includes a possibility or even expectation of reciprocation of help if a need occurs. But this possible future help does not need to be a condition for helping. If it is, the help is not altruistic. (The unintended consequences may be, however, the evolutionary reason for the tendency to have evolved.) If the second condition is not satisfied, the helping behaviour is a case of **behavioural pseudo-altruism**.

An important aspect of an altruistic trait is the context of behaving altruistically, or the range of altruistic behaviour. Philip Kitcher (1997) distinguishes four dimensions in altruism: intensity, pervasiveness, extent, and empathy. *Intensity* refers to how much weight the agent gives the good of the recipient, compared to their own good. There is pure egoism at one end, and "ultra-altruism" (taking only others' interests into account) at the other.[178] *Pervasiveness* refers to the variety of interactions in which the agent behaves altruistically. *Extent* is the extent of the individuals who are targeted by altruism. *Empathy*, in this context, is the ability to correctly read the recipient's needs or preferences. A fifth dimension, related to this, is *proactivity* of help; chimpanzees, for example, react only to signals of need, but bonobos (like humans) are proactive in helping (Melis 2018). Kitcher does not distinguish between psychological and behavioural altruism, and he uses the term "psychological altruism" in his discussion, but these dimensions may be applied to both: either they are properties of the altruistic behavioural tendencies or specify the contexts in which the empathetic concern arises.[179] From the evolutionary point of view, the behaviour is what matters.

Kitcher discusses differences between individuals in their tendency to be altruistic, but it would be more accurate to make these evaluations trait by trait. In fact, it is extremely important to qualify the traits in the right way for the evolutionary analysis. A particular

[178] This is similar to the distinction by Sober 1989 between different strengths of altruism that was discussed earlier.

[179] Note that although the tendency to be an altruist may be quantified on the level of a trait, the difference between altruistic and selfish behaviour remains a qualitative difference.

kind of helping behaviour may seem like evolutionary altruistic in isolation but turn to be part of a more general behavioural trait that, when all the occurrences and their contexts are taken into account, is fitness-increasing overall and there just happen to be some instances that would be fitness-decreasing in isolation. For example, helping behaviour might evolve because it is mostly directed to kin or reciprocating partners, but the selected behavioural inclinations result in systematic help of some other individuals too. There could be selection pressures to weed out the non-beneficial interaction, but if the psychological capacities to distinguish between the cases are not readily available, or if there would be other fitness consequences for doing so (for example, maintenance of the capacities or too much error in making the distinction), this does not lead into the evolution of a difference-maker, even if natural selection is the only factor considered.

6.3.2. Biology of Psychological Altruism

Behavioural altruism includes the idea that there is a psychological tendency to help without this help being an instrumental means to something else. This is the *function* of whatever psychological mechanism instantiates it. **Psychological altruism** is a specific kind of motivation mechanism – empathetic concern, as discussed previously. There are alternative psychological mechanisms that could underlie the same behavioural disposition and make both behavioural and agentive altruistic descriptions true. There could be various egoistic motivation mechanisms that accomplish the same function in most cases (such as becoming distressed from perceiving distress), or motivation mechanisms that do not consider the welfare or preferences of any party, such as making a moral judgement or executing an action script that may be, for example, an internalized social norm or just a habit. As Elliot Sober and David Sloan Wilson (Sober & Wilson 1998) point out, however, there is a significant causal difference between the different mechanisms: different interventions on the system have different effects. For example, if Beatrice's helping of Amos depends on

the distress that she experiences from seeing him in distress and there is no empathetic concern, medication that leaves Beatrice numb to distress would remove the motivation. With genuine psychological altruism, this would not happen. According to the pluralistic motivation theory (Batson 1991 & 2011; Kitcher 1997; Sober & Wilson 1998), there are multiple mechanisms, both empathetic concern and distress, but also moral and reflective sources for action.

Sober and Wilson (1998; Sober 1989) are careful to keep psychological altruism and what they call "apparent altruism" apart, precisely because of the different counterfactuals that the different mechanistic bases support (cf. Kitcher 1997), but they do not pay behavioural altruism (their "apparent altruism") the attention it needs. It is this tendency that is mostly relevant for fitness purposes, but also for other evaluations of someone being an altruist. Sober and Wilson rightly point out that the differences between the motivation mechanisms translate into differences in the *reliability* of the selected behaviour, which is fitness-relevant (see also Batson 2011). This is their classical evolutionary argument for the existence of psychological altruism, in a nutshell. They also point out, to undermine a counterargument, that there is no reason to believe that altruistic preferences would be less *obtainable* as the ultimate preferences, especially given that there are altruistic instrumental preferences anyway. However, the distinction between psychological and agentive descriptions makes a difference here. On the agentive level, there is indeed no conceivable difference between ultimate and instrumental preferences. But this does not hold on the psychological level. If the cognitively primitive organisms we descend from had a simple motivation system that worked through pain and pleasure only as proxies (that is, the motivation system is hedonistic), it would have been likely for us to evolve into beings whose altruistic tendencies were mediated through pleasure from helping and distress from others' distress.

As it happens, this scenario is not very likely to be true. According to Stephanie Preston and Frans de Waal (2002; Preston 2007; de Waal 2009), empathy at large is an evolutionarily very old and robust

motivation system, in which very primitive motivational and cognitive structures (that is, empathy and mimicry) are layered with cognitively more sophisticated layers of motivational and cognitive structures (as usually happens in evolution) (see also Batson 2011; Churchland 2011). In de Waal's "Russian doll model" (de Waal 2007), the innermost layer consists of capacities and tendencies to motor mimicry and emotional contagion, which are connected to each other. They are automatic state-matching processes and do not involve higher processing (as exemplified by contagious yawning).[180] The next layer includes coordination and shared goals on the cognitive (mimicry) side, and empathetic concern[181] and consolation on the empathy side. This involves understanding others' situations. The final layer has "true imitation" and emulation (that is, cognition involved in simulation) on the mimicry side and perspective-taking and targeted helping on the empathy side. There is also both behavioural and neuroscientific empirical evidence for the evolutionary earliness of empathy, and the popular theory is that its origin lies in parental care, but the system has been adapted to participate in other contexts for other functions[182] (de Waal 2009; Batson 2011; Churchland 2011). As Preston and de Waal also point out, there is an increasing self-other distinction going *higher* in the layers. The core of empathy is direct mimicry and contagion, bypassing any higher-level simulation.

The empirical evidence indicates that genuine psychological altruism almost certainly exists. This is not important here as such, since the details of the mechanism that produces altruistic behaviour are not crucial in the discussion after this section. It is, however, instructive to take a closer look at how Sober and Wilson (1998) present the argument, and at Steven Stich's (2007) criticism of it. Stich argues that a richer conceptualization on the psychological level allows for

[180] The processes on this level are also a good example of the enactive, non-representational part of cognition that was discussed earlier (e.g. Hutto & Myin 2017).
[181] De Waal calls this "sympathetic concern", but it is the same thing.
[182] The other functions include altruistic behaviour in other contexts, but altruism is not the only function of empathy. It also exists to reduce aggression.

alternative descriptions of preference structures that do not need to have ultimate altruistic preferences. It is enough that we have *sub-doxastic states* (Stich 1978; see also Fodor 1983; Stich 1990; Corey & Spelke 1996) that are representational but are not inferentially integrated, conscious, or changeable. These representations could be innate (in the sense I define in the next chapter) and selected to direct altruistic behaviour, and they could make an *instrumental* altruistic preference reliably present. For example, if the content of the fixed sub-doxastic state is "I will feel bad if I do not help my child," hedonistic motivation will reliably result in helping the child. This means that Sober and Wilson's evolutionary argument from reliability for psychological altruism does not work.[183]

The validity of the argument for the existence of altruism is not important for the main issue at hand, but there is another conclusion to be drawn from this critique. Within the distinction between psychological, behavioural, and agentive altruism, helping is both behavioural and agentive altruism regardless of the psychological mechanism involved. If we cannot distinguish between genuine psychological altruism and agentively and behaviourally altruistic psychological pseudo-altruism in any way that is connected to introspection, behavioural tendencies, or practically possible manipulations to the motivation structure of the individual, how important would a psychological altruism like this be? However, as Stich (2007) himself says, it is not clear what the hedonist and altruistic claims are. If the sub-doxastic state that Stich proposes were fixed and non-conscious and reliably produced altruistic preferences (albeit *strictly speaking* instrumental), should we not consider this a part of the empathetic-concern mechanism that *utilizes* a hedonistic system as its *subsystem*? What exactly would be the difference? Stich stipulates that the instrumental altruistic preference within this process is the only altruistic part of

[183] See also Rosas 2002 for criticism of the reliability argument. Nevertheless, as I said above, the empirical evidence for the existence of the empathetic system seems to be strong anyway (de Waal 2009; Batson 2011; Churchland 2011).

the process, but if the fixed sub-doxastic state is bound to this preference (and this is Stich's stipulation, without which the argument would not go through), this pair together constitutes a robust altruistic motivational structure that has no further motivation behind it. The only reasons to keep the two parts apart would be to interpret the "preference" in question as an agentive-level preference and therefore distinct from the representative states on the psychological level analysis. On the agentive level of analysis, however, the action guided by this system would be altruistic anyway.

But there are other, less speculative contexts in which failure to distinguish between the levels and the different notions of altruisms results in fallacy. For example, Joseph Butler (1726) famously argued against psychological hedonism by showing that internal hedonistic states are not the goals of actions. Elliot Sober (1992c; Sober & Wilson 1998) has pointed out that this argument confuses the different types of motivational factors for action: external and internal. Sober does not make the distinction in these terms, but this is the difference between psychological and agentive. Philip Kitcher (1997) seems to make the same mistake in discussing altruism. As mentioned, he does not distinguish between psychological and behavioural altruism. He argues for altruism by claiming that even the hedonistic theory presupposes altruism. In his example, a child is hungry and crying and his mother feeds him to get rid of the distress this causes in herself. But, Kitcher argues, the mother's distress presupposes that she is concerned for the child's wellbeing, and the distress is caused by this concern. Therefore, altruism exists in this situation. But which altruism? The distress is the motivation mechanism that the crying triggers. The action in this example would be psychologically selfish – but agentively and behaviourally altruistic. Again, what matters in helping is that there is an altruistic behavioural tendency that has a psychological implementation. This, in turn, is also behavioural altruism in the biological sense, as we have stipulated – it is about one person (mother) doing something to another person (child) that has the fitness-transferring property as a type of activity. **Evolutionary altruism**, on the other hand, is

strictly about real fitness consequences. A mother helping her child increases her own fitness.[184]

6.3.3. Kinds of Altruism

To sum up the chapter, this is how the main altruism concepts are related:

Seemingly altruistic behaviour: helping another individual (measured in goods that have fitness consequences).

- *Is helping voluntary and the intrinsic goal of the behaviour?* If yes, it is **genuine behavioural altruism**, if not, it is **behavioural pseudo-altruism**. If we are discussing an individual action in a human context, genuine altruism includes the identification of altruistic intention, which makes it **agentive altruism**. If the intended altruism is the ultimate goal of the action, it is **genuine agentive altruism**, but if it is an instrumental goal for something else that is self-serving, it is **agentive pseudo-altruism**. In a biological context, the behaviour needs to be a part of a *behavioural trait* that is altruistic as measured by its apparent fitness consequences. This is **quasi-intentional biological altruism**. A psychological *trait* is behaviourally altruistic if it is identified with a tendency to act altruistically.

[184] I will return to kin selection and inclusive fitness in the last part of the thesis, but it should already be noted that helping offspring should be considered an increase in *direct* fitness, not indirect. The core idea of fitness is not only to have offspring, but have them breeding further, which means that helping one's offspring to grow to adulthood is a part of direct fitness. (However, the child using resources that could be used to raise the next child is against the parents' evolutionary interests, which leads to parent–offspring conflict.)

- *Is helping guided by empathetic concern?* If yes, it is **genuine psychological altruism**. If the guiding mechanism is different but there is a psychological basis directive to behavioural altruism, it is **psychological pseudo-altruism**. Psychological altruism may be strong or weak depending on whether altruistic motivation can override self-regarding motivation.

- *Does helping increase another individual's fitness?* If yes, it is evolutionary altruism. If this increase is compensated for by increasing the fitness of the agent indirectly, it is **evolutionary pseudo-altruism**, but if not, it is **genuine evolutionary altruism**. Evolutionary altruism can be strong or weak depending on whether it decreases absolute fitness in the agent or only relative.

If apparently altruistic behaviour is manipulated by the recipient or it is conditional to some other benefit for the agent, it is behaviourally pseudo-altruistic. The condition may be, for example, reciprocal help in the future, or direct reciprocation with other goods.[185] Much of helping behaviour is probably conditional pro-sociality like this. But if the only (external) goal of the behaviour is the help, it is genuinely behaviourally altruistic. The altruism of the action can be measured in wellbeing, pleasure and pain, the accomplishment of the goals of the recipient's action, or his or her preferences. These are different measures that entail different things being altruistic under different measures, but the core of the concept is other-directedness. In human contexts, the other-directedness is related to the intentionality of the other-directedness. An individual act that is intentionally other-

[185] In animal contexts, things such as grooming in exchange for holding babies or sharing food in exchange for sex are examples, and these interactions can be modelled, for example, as biological markets, as trading of services and partner choice (Noë & Hammerstein 1994a, 1994b & 2016). The alpha male of a chimpanzee group who makes sure that the other individuals treat each other fairly in order to keep the subjects happy and stay in power (a better strategy to stay in power than bullying) is another example (de Waal 2013).

directed is agentive altruism, but traits do not have agentive descriptions. In biological contexts, there must be some analogy to intentionality, and the measure of altruism is in the transfer of goods that have fitness benefits. This is quasi-intentional biological altruism. Speaking of human behavioural traits, (human) behavioural altruism and (quasi-intentional) biological altruism are not the same but overlapping, and from the evolutionary point of view, the main target of explanation is the existence and function of behaviourally altruistic traits that are both intentionally other-directed and transfer goods that have fitness consequences. Evolutionary functionalist explanations for behaviour like this try to show that the apparent fitness transfer has a function that makes it directly or indirectly selfish in the evolutionary sense, nevertheless. Evolutionary selfishness and altruism are related to the actual fitness consequences.

The psychological mechanisms that guide altruistic behaviour are another object of explanation. However, the existence of genuine psychological altruism is not necessarily the most crucial issue in understanding altruistic behaviour. There is a variety of cognitive and motivation mechanisms that produce the behaviour together, and if we are interested in the basis of altruism, we should be interested in all these mechanisms in greater detail both on the neural and the psychological level. These mechanisms manifest in capacities and tendencies, and the mechanism that produces the behaviour in the given context is constituted by these parts and their orchestrated activities in the context.[186] Agentive altruism will not have evolutionary explanations of any kind – that would be a category mistake – and we should not let the agentive level guide our thinking too much about the functions of the psychological capacities either.

Following Marr (1982), I have distinguished three levels of analysis: the function, how this function is accomplished in what the mind

[186] As a reminder: although I refer to human psychology in mechanistic terms and this as such is controversial, the notion of mechanism I discussed previously is very liberal in what things can constitute parts and connections.

does (the algorithmic level), and how this (what mind does) is instantiated in neural processes. In the case of psychological altruism, the implementation is important, since altruistic tendencies are built upon more primitive operations (de Waal 2009; Batson 2011; Churchland 2011) – one could even say that not only are there no hedonistic constraints on psychological altruism, but *given the way the mammal brain works, there are empathetic constraints on purely hedonistic motivation systems*.[187] How psychological functions are implemented on the neural and algorithmic levels are two different *explananda* in the evolutionary explanation of altruism. But the "basics" of psychology, such as empathetic concern or emotional contagion, are also associated in a variety of behavioural traits and no behavioural trait can be identified with a specific psychological capacity. There may be a rare exception and we can always define a behavioural trait as everything a given psychological characteristic is involved with, as discussed in the first subchapter of this chapter, but this changes the topic from functionally defined behavioural traits to something else.

6.4. Individualism and Holism in Behavioural Traits

Abandoning individualism by dissecting the cognition and motivation mechanisms in greater detail opens the door for holism. This may sound paradoxical, but just as cognitive processes can be selected for their functionality in an individual's isolated behaviour, they can be selected for their role in social interaction. I will explicate this point next and define evolutionary individualism and holism in the proximate dimension. I will then argue for interactionist holism by discussing the evolution of reciprocal altruism. I will then make some remarks about human sociality and the possibility of collectivist holism.

[187] See Charkoff & Young 2014 for a review of the brain processes involved in prosocial behaviour.

6.4.1. Individualism and Holism in the Proximate Dimension

In an individualist evolutionary functionalist approach to behaviour, behavioural occurrences form traits according to the role played by the emerging behavioural patterns in the organism's life, understood through how the behaviour results in an achievement that makes individuals more adaptive in its environment. This was discussed with the example of the hornet-hunting behaviour of the honey buzzard. The behaviour is produced by capacities and tendencies that we can understand as a mechanism for that behaviour. The mechanistic basis also includes environmental factors, but they do not need to be part of the evolving system. However, sometimes they may be, as in niche construction processes (Odling-Smee, Laland & Feldman 2003).

What is of interest here is the possibility of cases where something that other individuals do is a necessary part of a mechanistic structure that contributes to the fitness-relevant achievement, and this doing of another individual is also a part of a mechanism that achieves something for this other individual. In other words, the evolving system requires interaction between individuals and the evolution of this behaviour involves the co-evolution of individual behavioural tendencies that take the behaviour of the partner into account. This makes the behavioural trait interactionist (it only emerges within an interaction) and the explanation holistic. The trait is interactive in two senses: the phenotype is interactive, but it is also constituted in the interaction between individual traits and group properties – that is, the makeup of the group regarding the trait. The trait's contribution to the fitness depends on the sum of the interactions the individual has in the various individuals they encounter within the group. Therefore, the trait is holistic, not just relational. If understanding the functionality of the behaviour requires the perspective of the whole group (such as in organized division of labour), the group forms a super-individual regarding this trait, in which case the trait is a trait of the group. The difference between the last two cases is the difference between interactionist holism and collectivist holism. In all cases, we are

interested in the function of behaviour, and in all cases some of the component parts of the behaviour are psychological processes. The difference between individualism, interactionist holism, and collectivist holism lies in whether the psychological processes participate in the production of behaviour that is functional and intelligible at the level of individuals, forms of interaction, or groups.

How realistic is it to assume such an interaction? What I describe above may sound like it requires intentional joint action of the kind that is studied in the action theory. This would make it something that could not plausibly be an object of the selection itself, although the capacity for such action might be. However, this is not the case if we abandon agentive explanation framework for behaviour, and approach behaviour as produced by cognitive capacities and tendencies, as I have argued.[188] I have already discussed the dual process theories

[188] Dan Sperber (1997) has argued for sub-individual cognitivism against individualism in social sciences based on greater realism and explanatory depth. The distinction between the "agentive" level and "psychological" level that I used earlier coincides with this distinction between "individual" and "cognitive" in some respects. However, my "psychological" is a general category of relevant sub-personal causal processes that participate in producing behaviour, regardless of the level of abstraction in description, and some of these might not be "cognitive", strictly speaking. Sperber criticizes traditional individualism, which he calls "strongly individualistic" (the explanatory factors are individual-level properties) and "weakly cognitivist" (the role of cognition is trivial), of being a non-realist and non-naturalist approach to human behaviour in the guise of being more realistic. He calls for a position that is "weakly individualist" and "strongly cognitivist" instead, in which explanations refer to infra-individualistic cognitive processes and are naturalistic and mechanistic in the sense that it takes cognitive science and biology, for example, more seriously. This also entails that cognitive factors are causes among other types of causes, not reason explanations sui generis. The rationale of behavioural economics (for example, Kahneman & Tversky 1979; Hogarth & Refer 1987; Kahneman 2003; Diamond & Vartiainen 2012) is to do exactly this. Jaakko Kuorikoski and Petri Ylikoski (2008) have criticized the speciality of individual agency in explanation – or intentional fundamentalism, as they call

of cognition and non-conscious behavioural guidance. There is strong psychological evidence suggesting that we react automatically to some cues in the environment, including the social environment, without deliberative processes, in ways that affect our social behaviour; some of our attitudes in social situations are non-conscious, and we may "tune in" with others automatically through direct emotional contagion and become motivated by the perceived needs of others, as discussed in earlier chapters (see Nisbett & Wilson 1977; Bargh & Chartland 1999; Haidt 2001; Preston 2007; Uhlmann et al 2008; de Waal 2009; Batson 2011). If these capacities and tendencies are selected, they may have been selected for interaction that is beneficial to participate in.

If we break the proximate mechanisms of behaviour into finer parts than individuals and explain the occurrences of social behaviour as the results of (probably multiple) cognitive processes (in a specific context), we can cluster these interactive processes into traits at any higher level. It is still sensible to talk about individual behavioural traits, for example. The psychological (or cognitivist) perspective does not exclude this; it only implies that the individual (or agentive) level has an internal causal structure. In the same way, we can take a context of social interaction and examine the behaviour of this "system" of individuals and how sub-personal psychological processes form a mechanism that guides the behaviour of this system.

We can define individualism and holism as follows:

> Behaviour is **individualistic** if it can be adequately described in terms of individualistic goals and/or behavioural dispositions and mechanisms regarding its adaptive function. This adaptivity can be fully understood in terms of the consequences of the individual behaviour that can be fully understood in terms of the individual aiming at something, whether this is an intentional goal

---

it – for similar reasons. These are arguments about explanation in social sciences, but they are also additional arguments for the importance of sub-personal processes in the causal explanation of human behaviour in general.

or something that we can attribute as the evolutionary function of the behaviour only.

Behaviour is **interactionally holistic** if it is produced through the interaction of two or more individuals such that the behavioural dispositions of participating individuals jointly produce the adaptive achievement. The individual motivational mechanisms constitute a social mechanism that is responsible for the adaptive behaviour and the individual behaviour is instrumental to this, regardless of what the individuals themselves would state their goals to be.

Behaviour is **collectively holistic** if the underlying mechanisms are fully socially contextualized in their function. That is, the proper function of the behavioural response to a social stimulus, for explanatory purposes, is on the collective level, not on the level of goal-oriented individual agents. The behavioural traits are traits of a collective (a group) and the adaptivity is on the group level.

I will now further discuss interactionist traits through an example of reciprocal altruism. I will return to the collective traits after that.

6.4.2. Reciprocal Altruism

The evolutionary basis for some helping behaviour is future reciprocation by the recipient. As previously discussed, the evolution of helping that is paired with reciprocation can be modelled as a tit-for-tat (TFT) strategy. In an open-ended Prisoner's Dilemma game, a player of TFT strategy cooperates the first time and then mirrors the strategy of the co-player. This strategy does worse than pure selfish (S) when comparing two individuals, but in a population of selfish and TFT-

players, it is likely to emerge as the winning strategy.[189] This strategy reflects the *logic* of how reciprocity can promote helping behaviour, not necessarily the patterns of behaviour. A few clarifications related to this are in order.

First, given these are evolutionary games and strategies, selfishness and altruism are measured in fitness but not in the overall fitness consequences for the individuals participating – the altruism the evolution of which we are interested in here is behavioural altruism as defined above, not genuine evolutionary altruism. Conflating the two would lead to the *averaging fallacy*, which fails to distinguish the trait's fitness effect in the behavioural context and its net effect over the lifetime (see Sober & Wilson 1998; Okasha 2006); I will come back to this later. Second, the aims of the interacting individuals are not relevant as such. The agentive-level aims may be altruistic or not, and the psychological motivation may be altruistic or not. Third, the logic of TFT tells us why helping is beneficial if it is reliably enough targeted at those individuals who eventually help back, and the key element in this is to react to the individuals according to what they are doing. We do not need actual turns of a game or anything like that, for example, and the actions that count as helping do not need to be clear-cut.[190]

---

[189] If there are pure altruists in the population, they do worse than TFT if there are selfish players, too – otherwise there is no difference between pure altruists and TFT's. Whether TFT or the selfish strategy emerges as dominant depends on chance.

[190] One problem in interpreting models is that they do not only idealize the processes that they model, but they also abstract from their target system. If the model works, it captures some of the relevant behaviour of the system, and in the case of evolutionary game-theoretical models, I take it that they capture some relevant causal relations. They do not, however, specify what entities, properties, structures, interactions, or mechanisms instantiate these causal dynamics. A model does not need to be realistic in this sense to capture the logic of how traits get selected, for example. Conversely, just because a model works, it does not mean that there must be discretely identifiable counterparts to its key elements in the system. Assuming this would be a reification fallacy of the misplaced concreteness kind. I will not discuss the nature of

There are also other ways than just reacting to past experiences to couple the help with those who reciprocate: observing interactions with others (which can be modelled as *observer-TFT;* Pollock & Dugatikin 1992), the attitudes of others (that are based on experiences; Castro *et al* 1998), or reputation (Ohtsuski & Iwasa 2004). All of these lead into partner choice (which is **direct reciprocity** instead of reciprocal altruism; see Noë & Hammerstein 1994a, 1994b & 2016; Noë 2006; Noë & Voelkl 2013) and punishing the defectors (**strong reciprocity** of Herbert Gintis; 2000a & 2006). All these additions make the behaviour both more realistic in human contexts and stronger against selfishness.[191]

There are many ways to implement the evolutionary logic of simple TFT – or rather, there are several combinations of psychological configurations and behavioural interactions that can be explained as a TFT mechanism. The most straightforward psychological implementation is to have a directly conditional attitude. That is, Beatrice only helps Amos with the understanding that Amos will help her in

---

models in this thesis on top of everything else, but I assume that above said is non-controversial.

[191] The border between reciprocal altruism and directly cooperative reciprocity is not always clear. For example, if we look at the kinds of interactions instead of individual behavioural disposition, as I am suggesting and has been suggested (for somewhat different reasons) by Patrick Forbes and Rory Smead (2015), the standard evolutionary classification of behaviour falls apart. Second, if we are interested in an individual's behavioural dispositions or psychology, the focus must be their overall behaviour and its consequences over their lifetime, which means that the same individual traits can participate in different kinds of interactions in this regard. There are two differences between reciprocal altruism and direct cooperative reciprocation: the timespan between the reciprocation, which is never immediate and can be overlapping in complex interactions where the benefits of action are spread over time; and whether the "enforcing activity" is partner choice or partner control. The difference between partner control and partner choice is not clear-cut in long-term relationships, which may include both, and the underlying form of interaction does not change regardless of whether the partner is chosen or controlled.

the future. If she did not believe he would, she would not help him. A few remarks are needed again. If this psychological basis leads to systematic behavioural altruism, it is a psychological mechanism for behavioural altruism. If Beatrice has a reliable way to know that Amos will reciprocate and she chooses to help him because of this, the discrimination (or partner choice) makes the behaviour a case of direct reciprocity in the evolutionary sense. If she does not know this and takes a leap of faith, it is reciprocal altruism. That is, this is the distinction between direct reciprocity and reciprocal altruism as it manifests in particular choices. However, for actual evolutionary mechanisms to exist, the psychological capacities and the related behavioural tendencies must be life-long tendencies. Furthermore, they need to be distributed across the population. The same psychological makeup could be selected by biological markets (direct reciprocity) and as a TFT strategy at the same time, if choosing the partner to interact with is easier at some times and more difficult at others, and this varies by the individual with the same psychological makeup. The psychology of reciprocity or reciprocal altruism like this cannot be altruistic (in the psychological sense) in any case, since helping is conditional by definition – it is not motivated by the need of the other but the projected future utility. Given that this conditionality is directed at external conditions, its instances are not agentive altruism either. This is important because if the parties all realize the conditionality, they understand the expected help to be conditional and this cannot fail to affect the interactions between the individuals.

There are alternatives to this *homo economicus* style of reciprocal altruism – for instance, *friendship* (see Hruschka & Heinrich 2006). One way to think about friendship is coalition-forming. Some of the literature on cooperation stresses control as a key component to rule out free riders (see Gintis 2000a, 2000b & 2006; Bowles & Gintis 2011) but partner choice may be a more efficient way to guarantee collaboration (Noë 2006; Sterelny 2012; Noë & Voelkl 2013). Keeping free riders out of coalitions for collaboration (such as hunting) could involve not only retribution and punishment but also commitment, signalling this

commitment in costly ways, and the trust this engenders. This has lately become a more common way to approach the foundations of group living. Coalition-building for things like hunting or power politics is, however, only one aspect of human social life and bonding between individuals. Other aspects include things like communal childcare (see for example Narvaez *et al* 2014), learning and teaching skills that do not necessarily involve only family members (Boyd & Richerson 2005; Sterelny 2014), sharing information (including "gossiping", which may play an important social function in trust and coalition building), forming peer groups for various tasks, including learning social skills (see Narvaez *et al* 2014), and a multitude of more mundane everyday tasks that require smaller-scale collaboration. Not all non-kin relationships need to include high risks, either. There may be a multitude of evolutionary reasons for friendship to evolve and a multitude of psychological-level instantiations. One obvious one is compassion, an affective bond that creates empathetic concern between friends. As discussed above, empathetic concern probably evolved for parental care at a very early evolutionary stage (see Churchland 2001; Preston & de Waal 2002) and once the mechanism existed, it could be adopted for other uses, such as romantic bonding (it being one of the psychological foundations for romantic love; see Fisher 2016), or friendship.

The details of the evolutionary history of friendship or its psychology are not important here as such but building bonds with compassion is a crucial part of human psychology (see Hoyyat & Moyer 2017 for the psychology of friendship and development of friendships). Now suppose Beatrice and Amos are friends, there is an empathetic concern between them, and this is why Beatrice is automatically motivated to save Amos. The help between them is not conditional on a predicted future. In fact, Beatrice may very well know that Amos will never be able to reciprocate to quite the same extent. Beatrice's help is guided by pure psychological altruism. This behaviour, and the psychological altruism as its basis, could still be part of an overall psychological and behavioural tendency that fulfils the

conditions of reciprocal altruism on the scale of lifetime. Developing a bond that includes genuine psychological altruism is the part of psychology that produces altruistic behaviour. The other part of TFT is to stop behaving in this way if there is a reason for it; there needs to be a capacity to disconnect the bond to avoid being a sucker. This can be implemented through negative feelings. For example, if Beatrice has helped Amos on several occasions and Amos is consistently reluctant to help her when she needs help, this may loosen the bond on Beatrice's side or even make her feel used, bitter, and angry, and cause her to end the friendship. On the other hand, signalling compassion and affection with things like mutual care and attention, even if they are not costly, may be required to maintain the understanding of the nature of the relationship. The disposition to end a friendship in case of non-reciprocity does not mean that there is any conditionality of empathy at the moment of help. The form of reciprocal altruism is an evolutionary formula that describes certain properties of interaction on the behavioural level that can be abstracted into a TFT strategy over time. There can be selection for this form in interaction – for example, for friendship. There is also a selection to be able to participate in friendship relations, and for the psychological characteristics that are needed for this.

Friendship-related behaviour and psychology are about relationships between individuals. Because of our limited capacities for bookkeeping, selectiveness about the partners of interaction and forming long-term relationships with them might have selective advantages (Hruschka & Henrich 2006), but a form of interaction in which people collaborate can be about specific contexts or forms of interaction. Whether the evolving interaction is an individual behavioural trait (that is, related to a specific achievement and specific type of interaction) or a person-directed form of relationship, and whether the explanatory mechanism is direct reciprocity (on the evolutionary rather than the psychological sense) or a TFT-type of selection logic, there are three different elements at play: behaviour, psychology, and the folk-psychological interpretation.

First, there is the form of behaviour as a behavioural trait. We can identify two different behavioural traits: the individual behavioural patterns and the interactive behavioural patterns. To understand the adaptive significance of reciprocal altruism, we need to pay attention to the interactive trait. The adaptive significance of participating in reciprocity (generally speaking) comes from the benefits in return, and this takes place only in interaction with suitable partners. Elliot Sober and David Sloan Wilson (1998) have argued that this shows that the evolution of reciprocal altruism as a TFT strategy is a form of multilevel selection, despite emerging within an apparently individual-based model. The idea is this. In addition to the competition between individuals within the pairs of interactors (for example, between selfish (S), altruist (A) and TFT types), there is a competition between groups that consist of the interacting individuals: TFT/TFT, A/A and A/TFT pairs (which have group benefits) do better than S/S, S/A, or S/TFT pairs (which do not).[192] I will return later to whether this is the case and what is a "group" for multilevel selection purposes. However, at minimum, there is an adaptively significant difference between the different types of interaction; whether these qualify as *trait groups*, they involve *group traits*. The benefits of participating in a beneficial interaction depend on the partner, but they do not necessarily depend on the individual variant: TFT/TFT, A/A and A/TFT interactions are identical. In a population with several possible types of interaction, there is a competition between these types of interactions, just like there is a competition between individually implemented traits. Interactive traits (or interactive phenotypes) depend on the individual behavioural dispositions, which are constitutive parts of the interactive traits. These traits can have two adaptive functions: to enable participation in the interactive traits and whatever they achieve for the individual within the interaction.

[192] The selfish interactor receives the benefits from altruism of the altruism in the S/A pair, but this is not a case of a group benefit.

Furthermore, the traits cannot be singular occasions of behaviour from the evolutionary point of view; they must be behavioural patterns over the lifetime of the individual. In principle, an individual behavioural trait could be a type of behavioural response, the set of alternative responses that are available to the individual in the same context, or all possible alternative behavioural responses that exist in the population. In the first case, the same individual behaving altruistically towards one individual and not another would be two different traits. I have already argued for why this is not plausible, but to rephrase this with the case under discussion: we would not have behavioural traits in the TFT form at all, just individuals with both selfish and altruistic behavioural tendencies, different traits being triggered in different contexts. This may be a sound way to conceptualize traits for some purposes, but it would lose the idea of behavioural strategies as traits and the connectedness between the behavioural responses. We need to include the entire behavioural repertoire of the individual in a given behavioural context as the trait, and the choice between the available responses is a part of the trait. This also means that the alternative repertoires of different individuals are alternatives of the same trait: selfishness, altruism, and different mixed repertoires with different criteria for the choice of behaviour. The interactive traits, on the other hand, are constituted by interactions. The interactive phenotype of collaboration is constituted by the behaviour of collaborators, regardless of whether they "play" A or TFT as individuals (that is, which other types of interactions the individuals participate in). The interactionist reciprocity trait is a trait of the pair reciprocating.

A behavioural trait requires a psychological capacity for the behaviour. This includes a motivational mechanism that motivates (in this case) altruistic behaviour. The relevant sense of altruism to be motivated is behavioural altruism.[193] The details of a motivational

[193] To be more specific, behavioural altruism in the biological sense. We can conflate biological behavioural altruism and behavioural altruism measured in some psychologically accessible way (that is, well-being, pleasure, wants,

mechanism do not matter for this: the motivation may be empathetic concern (genuine psychological altruism) or distress from perceiving distress (a psychologically egoistic motivation for agentive altruism). It may also be something that does not lie on the egoism–altruism dimension at all, such as a *behavioural script* (in the sense of Tomkins 1987; an automatized behavioural pattern that may be, for example, an internalized social norm), or a moral judgment or a principle.[194] What matters, however, is whether it is altruistic under the agentive description: is the intended achievement of the behaviour to help (or cooperate)? And if it is, is it the ultimate goal of the action (genuine agentive altruism) or an instrument for achieving something else (agentive pseudo-altruism)? If it is instrumental, what for? Another psychological capacity needed for reciprocal altruism (in any of its form) is to recognize whether the partner is reciprocating or not. Recognizing the psychological mechanism is not important, but recognizing intentionality is.[195] In sum: reciprocally altruistic interactive traits include the individual psychological basis (including a motivational mechanism that may be psychologically altruistic or not, folk-psychological capacities, and the ability to both form and break a bond); behaviourally altruistic individual behavioural tendencies that can be

---

or some such) as discussed above, although in real world cases there will be some discrepancy.

[194] There are different possible motivational mechanisms for producing similar behaviour. There may be differences between them, however, in other aspects that I discussed in the previous chapter: in their extent, strength, and reliability. Some of these differences may have adaptive consequences, which means that they are not evolutionarily exchangeable. However, motivational pluralism is the likeliest psychological configuration for both empirical and evolutionary reasons (see Kitcher 1997; Sober & Wilson 1998; Batson 2011).

[195] This is one reason why distinguishing agentive and psychological is relevant. Knowing the psychological basis for wanting to help may be relevant to knowing how robust the tendency is, but this is not the same as asking whether the intention was *really* to help. Again, folk psychology may not make this distinction but conceptualizes the basis as "more ultimate reason" or something on those lines.

recognized as agentively altruistic; and the interactive phenotypes that the individuals form in various combinations. I will draw some conclusions from this and generalize them now.

6.4.3. Psychological and Behavioural Traits in Interaction

When a behavioural trait is selected for, the psychological traits that make it possible are selected. Psychological traits are selected for the behavioural consequences they have. Any given behavioural trait involves several psychological capacities and behavioural tendencies, and any given psychological trait may participate in several behavioural traits. This means that the functional structure of psychology and the functional structure of behaviour are not isomorphic. There is co-evolution of psychology and behaviour where both are relevant to each other. There may also be other factors involved; I return to this in the next sub-chapter. This entails that behavioural traits must be treated as separate traits for analytical and explanatory purposes: the adaptivity of the behavioural traits in which the psychological capacity participates is what determines the adaptivity of the psychology. Behavioural traits are bound to psychology being caused by them (jointly with the environment) in two different ways: behavioural traits using the same capacities (having a shared causal component) ties the behaviours together proximately and evolutionarily. However, if the psychological capacities are flexible and several non-specialized capacities participate in the behavioural traits, they may have a too limited life of their own to be evolutionarily significant.

The facts about the causal connections and behavioural patterns enable several sensible ways to break the behaviour into traits, but under the evolutionary functionalist approach, the perspective is an adaptive achievement that is produced by repeatable patterns of behaviour. The patterns may be abstractions, as discussed. The psychological capacities and motivational mechanisms that are needed for the achievement constitute the mechanism for it. If there are no reasons to restrict the analysis to individual traits, the analysis can in

principle be used to describe interactive traits as well. I have given some reasons for why individualism cannot simply be assumed and why (folk-psychological) intuitions about the primacy of the individual are misleading. For a trait to be an interactive trait, its achievement must be caused by the behaviour of multiple individuals in a way that the individual psychologies and behavioural dispositions form the mechanism to do so.

New systemic properties can emerge simply as a result of individual properties and their relations (see Wimsatt 2007) and there is no reason to consider the structurally emergent properties necessarily "holistic" – they can be part of an individualistic understanding of behaviour (see Ylikoski 2017). There must be some further criterion for what counts as a holistic approach and what difference it makes. I will discuss collectivist holism in the next subchapter; but for an interactionist holistic perspective within the evolutionary functionalist framework, the criterion is adaptivity caused by interaction. I have been vague about whom the behaviour should be functional for, the individuals or a group they are constituting, since this is a further issue about individual and multilevel selection. It could be either or both, but for now we can concentrate on individuals. Within this framework, all interactive behaviour that has emergent consequences that are not simply aggregates and make the individuals more adaptive because these emergent consequences, can be considered as interactive traits. This may be too permissive, however. It is no more permissive than any other adaptive functionalist analysis of an individual behavioural trait if the analytical perspective is simply the current use, but even then, there must be a reason for choosing the holistic evolutionary functionalist perspective other than just because we can. As I have argued before, mere current use analysis may also be of use (ecological function, analysis of evolutionary potential and maintenance of the structure, and so on). Things become more interesting, however, if the perspective is explanatory, whether historical or ahistorical.

If the emergent structural properties are the source of adaptivity, these properties may be the reason for which they are or were selected.

There is a holistic mechanism, but this mechanism is evolutionary, not proximate. Consider the following simplifying thought experiment. Suppose there is a species of primates that eat fruit that grows in bundles on the upper branches of a species of tall tree. Getting to the fruits is somewhat laborious and dangerous because of birds of prey, but once up there, it is no more trouble to drop the whole bundle down than it is to pick just a couple of fruits that the individual will eat. Consequently, other individuals get their share. This is an evolutionarily altruistic trait that can perhaps evolve through group selection. It is also a trait that involves multiple individuals in terms of adaptive benefits. However, there is nothing in the behaviour itself that requires the activity of several individuals and the behaviour itself does not make the dropper fitter. If group selection is operational, all the members of the group do better than members of the group of selfish primates that only eat the fruit themselves after reaching them, which means that there is a mechanism that makes altruism beneficial for the altruists, but this is a selection mechanism working over evolutionary time.

Now contrast this with reciprocal traits. By definition, an altruistic act by an individual is beneficial for the individual themselves only through reciprocal altruism. The adaptive achievement is the mutual aid and the mechanism for this is the participation of both parties in the interaction. In this case, the connection is in the proximate dimension. Furthermore, this is not a case of only putting individuals together and seeing what emerges from their interaction. Using a game theoretical model makes it look like that, but these models do not show the details of the causal structure of the interactions, as I have argued above. They show fitness effects but not what makes them. The phenotypes are not static; the same individual has the same behavioural dispositions in all interactions and makes a choice depending on whom they are dealing with. In other words, the phenotype itself depends on the partner. As I have mentioned earlier, biologists Allen Moore, Edmund Brodie and Jason Wolf call such social traits

interacting phenotypes (Moore *et al* 1997; Wolf *et al* 1999).[196] The idea is that certain behavioural traits *require* a social trigger and emerge only in the social context. Furthermore, the behaviour of the partner is a part of the social selection environment, and the social evolution has an interactive dynamic which they are able to model using indirect fitness similar to that used in kin selection models. The same logic can be expanded from pairs to networks (Formica *et al* 2017; Montiglio *et al* 2017).[197] These models are ostensibly individualistic in the evolutionary dimension in the same way as game-theoretical and kin selection models. The multilevel-selectionist interpretation of reciprocal altruism in game theory (Sober & Wilson 1998) could be applied here, too, if it works in the first case. It seems, however, that there are two different ways to link the individuals and their traits with adaptivity: over evolutionary time through group selection and directly through connectedness of the traits.[198] I will return to this issue in the final chapter.

The point now is that individual behavioural traits are bound together through their adaptive function. The individual *interacting phenotypes* form a multi-individual *interactive phenotype*. Just as we can understand some individual behavioural occurrences by an individual only as a part of an individual behavioural *trait* that emerges during their lifetime, we can understand some other individual behavioural occurrences by an individual only as part of an interactive

---

[196] My discussion does not depend on this specific theoretical work or how these models work. However, it is noteworthy that the idea of phenotypes that only exist in social interaction and hence are not properties of the individual alone, and that some behavioural traits are evolutionarily intelligible only when approached as a part of the interaction, has emerged within individualist paradigm of theoretical evolutionary biology too, not only among the proponents of group selection.

[197] I am only discussing interactive pairs for simplicity, but presumably my discussion could be similarly generalized to groups of multiple interactors.

[198] There is also a third way: the link in the developmental dimension through dependency in reproduction of the trait. I will return to this in the next chapter.

phenotype. There is one crucial difference, however. The interactive traits are created by the interacting individuals as individual characteristics of those particular pairs. A reciprocal altruist takes part in various relationships and each relationship has its fitness effects depending on the reciprocity nature of that particular relationship. The underlying psychology, which is the most individualistic part of the interaction, stays the same. The psychological characteristics of a person do not necessarily stay the same throughout their lifetime, but we can assume for the sake of simplicity that an individual has a more or less fixed psychology, regarding the interaction dispositions. The individual relationships and context vary. Say for example that Beatrice is a person who tends to be helpful and to intensify this helpfulness as the relationship is "tested" through time (this does not need to be conscious in any way), but also to distance herself after disappointing experiences. This complex set of interacting psychological dispositions results in a pattern of behaviour over her lifetime that qualifies as a TFT strategy over evolutionary time. This same evolved psychology is present in all individual relationships, but if Amos reciprocates and someone else, say Curt, does not, the relationships will develop in different directions with different interactive characteristics. The same applies if the tendencies are only about a particular kind of interaction. The adaptive function of the psychological capacity to participate in reciprocal relationships comes from participation in those forms of interaction that are beneficial. It is not the individual behaviour regardless of the context. There are two evolutionary mechanisms: the competition between forms of interaction (that decide the fitness consequences) and the competition between psychological makeups that enable participation in the beneficial forms of interaction and prevent entering or staying in dysfunctional ones.

If there is selective competition between individual behavioural traits, there is selection for psychological dispositions that cause this behaviour. If there is selective competition between forms of interaction, there is selection for psychological dispositions to participate in the interaction: to cause the behaviour under certain conditions. To

understand the psychology of interactions from the evolutionary point of view, it is necessary to understand how it fits with the different contexts – what is the function within the mechanism that produces interactive traits. This makes a difference between individualist and interactionist approaches in understanding the proximate-level causal processes that are involved in social evolution. Consequently, this difference is significant for the use of evolutionary considerations as a guideline to understand human behaviour and psychology. The right way to do evolutionary psychology, for example, is not necessarily the individualistic way. Furthermore, since the selections for psychology and for behaviour are different, the selfishness and altruism of related psychological and behavioural traits do not necessarily go hand in hand either. Remember that the psychological traits are selected for their lifetime fitness effects in all the interactions they participate in. In an environment of reciprocal altruists, being another reciprocal altruist is an evolutionarily more selfish strategy than being behaviourally selfish. Having the psychological makeup of a reciprocal altruist may include having a psychologically altruistic motivational mechanism that leads to behavioural altruism (measured in fitness, too) under right circumstances, but it is an evolutionarily selfish psychological configuration under these circumstances. This is not the case with pure behavioural altruists – this is not an evolutionarily selfish strategy even under those same circumstances.

I will come back to the group selection issue and why treating TFT strategy as selfish is not an averaging fallacy (contra Sober & Wilson 1998 and Okasha 2006) while treating a pure behavioural altruism as being evolutionarily selfish would be. Sober and Wilson and others are, however, correct in highlighting the importance of focusing on the pairs of interaction (as groups) instead of individual traits. I will also agree that the evolution of altruistic or reciprocal tendencies as a TFT strategy is a form of group selection, but a different form. For now, the conclusion from the discussion so far is that from an evolutionary functionalist point of view, human social behaviour involves supra-individual interactive traits, the individual behaviour involved

should be understood from this point of view (as component parts of something more holistic), and the underlying psychology should be approached as selected for participation in such holistically understood forms of interaction. Agentive descriptions, on the other hand, may play a role within the practices in which the interactive traits emerge, but these descriptions should not be considered as descriptions of what the behaviour is about (in the wider context), a part of the description of a behavioural trait, or as references to the relevant psychological mechanisms.

6.4.4. Evolution, Sociality, and Collectives

The main issues in the evolution of human social behaviour are to explain why we have prosocial tendencies in the first place and why they take the forms that they do. The approaches to these issues have been mostly individualistic in the proximate dimension: the interest has been in individual behaviour and its evolved psychology. Some of the approaches have been holistic in the evolutionary dimension, using group selection explicitly, and some have used models that can perhaps be interpreted as being implicitly holistic. Some of these approaches attribute traits to whole groups. For example, some of the advocates of using group selection in human context have treated human groups as holistic entities in the sense that I have called collectivist, and attributed traits (and accompanying fitness consequence) to whole groups. I have argued so far that psychological and behavioural traits should be kept substantially apart in understanding human sociality and that this entails a possibility of holistic behavioural traits, and I have argued for the interactionist version of such traits, but the collectivist version is possible within this framework too. However, these are two different ways for something to be a trait above the individual level.

When group selection fell out of fashion the first time around, the evolutionary approaches to social behaviour became individualistic in other senses too (as discussed in chapter 4). The return of group

selection models changed this again to the degree that group selection was accepted. Cultural group selection (Boyd & Richerson 1985 & 2005; Wilson 2002) added a group perspective to anthropology as a way to understand the differences between cultural groups. Group selection also gave a new perspective on the *major transitions in evolution,* the transformations in biological organization where entities of one level unite into entities of another level, resulting in change of levels in where the biological individuality takes place (such as the origin of multicellular organisms; Maynard Smith & Szathmáry 1995; Calcott & Sterenly 2011; Bouchard & Huneman 2013). As David Sloan Wilson put it, the transition from "groups *of* individuals to groups *as* individuals" requires group selection, since it is difficult to see how a group of individuals could suddenly become an individual of higher level without there being evolution of "group-level" properties and their harmonization that constitute the unity and individuality of the group before the transition already (Wilson 2002: 17). In addition to cooperation between individuals, the division of labour is an obvious example of an inherently group-level trait (see Hamilton & Fewell 2013).

There are critics of this thinking, however. As already mentioned, the essential problem of group selection is that individual-level selection is too strong for group-level selection to establish group-level adaptations. This is why collectivist or *superorganismal* traits are rare. There seem to be some examples of group selection, but critics such as Andy Gardner and Alan Grafen argue that since they are all cases of either kin selection or traits that have evolved to suppress within-group conflict, they are not really cases of group selection (or group adaptation) but of individual selection (for example, Gardner & Grafen 2009; Gardner 2013b). Gardner and Grafen consider group selection only where there are both competition between groups and lack of competition within the groups, since this is where the individual level does not trump the group level (see also Brandon 1982 & 1988; Maynard Smith 1988). In other words, they consider groups as a unity for adaptation only when they have already become individuals; for example, Charles Goodnight, an evolutionary biologist

working in the multilevel selection framework instead, defines individuality as the *lowest* level of biological organisation that evolves adaptations (Goodnight 2013). This also means, for them, that there are only individual level adaptations.

I will not go deeper into the issues of individuality or the major transitions in general. I am only interested in the human context here. However, the disagreement over group selection is not only about what counts as group selection and an empirical issue of whether it exists. There is also an issue about the level of adaptation of the adaptive social behaviour, which I consider to be a separate question. My interactionist holistic account of social adaptations is a suggestion for a middle ground: regardless of whether the selective forces that sociality work on the individual level or on multiple levels, adaptations of interaction emerge. They may be adaptations *of* individual behavioural dispositions *for* emerging interactions, but as I have shown above, these interactions are the evolving traits that make the difference for the fitness consequences for the individuals participating in them. Psychological adaptations are for beneficial behavioural traits that are beneficial depending on an achievement that has fitness consequences for the entities that bear fitness. This entails selection for the parts of the mechanism that produces the achievement, where mechanisms are abstract forms of interaction, the connections between parts are causal dependencies, and the parts of the mechanism are sub-personal psychological states and environmental factors. If the interactive behavioural trait requires orchestration such as responsiveness to other individuals' behaviour (the example of reciprocal altruism) or division of labour, these traits cannot be reduced to individualistic traits, the relations between individuals, and the emerging patterns of social interaction only. However, they *also* involve individualistic psychological selection, and, if individual selectionism is correct, these individuals are the beneficiaries of the fitness consequences that matter. On the other hand, if the evolutionary processes leading to reciprocal altruism, division of labour, and the like are multi-level processes, these traits may be conceptualized as group-level

adaptations if we consider the interacting individuals to be groups. This is partly a matter of defining a group – for example, are the trait groups of Sober and Wilson really groups, or should we go with the "broad sense individualist" interpretation (Sterelny 1996a; see also Kerr & Godfrey-Smith 2002)? – but there are also substantial differences between different kinds of groups from the evolutionary point of view, if we are liberal about what are groups. The third option for interactive traits is to be traits that also evolve in the interaction between individual- and group-level selection, as products of multilevel selection specifically. I will later return to why these distinctions matter and distinguish different notions of group selection.

The *levels* of selection (as the level of biological organization at which natural selection operates) and the *units* of selection (the entities that do the adapting, or the level of adaptation) are usually thought to be the same, usually implicitly, sometimes explicitly (Okasha 2006), with notable expectations such as Robert Brandon (1982 & 1988) and Elizabeth Lloyd (2001). I will suggest later that not only are these separate questions, but that both issues have two different senses in which the units or levels of selection could be group-level. It is not a contested issue that if the group-level properties and selection processes mask the individual level and the groups become superorganism, and the group fitness differences screen off individual fitness differences (Brandon 1988), this is a case of group selection. It is just as clear that group selection involving multilevel selection is different from this (see Sober 1992d). However, what I have implicitly argued for in this part of the dissertation, and am about to make explicit, is that behavioural adaptations (or adaptively functional robust patterns of behaviour) can exist on an intermediate level between the individual and group levels. Even if individualism about *selection* is correct, an *adaptation* can be interactive, such as reciprocal altruism.

Furthermore, there are two different, equally important evolutionary notions of group that may be needed to analyse even the same cases. One way to understand a group is as the collective of individuals who inhabit the same region and form a social group, a collective

(such as a tribe), which includes all the individuals. Another way is to define a group as the network (or even a pair only) of individuals who interact (and maybe choose to interact) which each other in respect some behavioural trait (that is, Sober and Wilson's trait groups). I have discussed interactive traits as traits that bind individuals together as groups in the latter way. If this facilitates a type of group selection, the group may still extend to be the whole collective. But it may also be that there is competition between interactive traits (and therefore trait groups) within the collective, and group selection between collectives, which may involve differences between the collective groups regarding the frequencies in which they have different interactive traits, for example. I will return this in the last chapter.

Pure group-level adaptations would be adaptations of the group as a collective, as a superorganism, where individuals (who still do the concrete behaving) are interchangeable with other individuals. This interchangeability may be absolute (any individual will do) or according to their role in the division of labour (such as the "castes" of social insects). The evolutionary functional perspective applies here the same way as with interactionist holism: the object of analysis may be current utility, historical explanatory functionality, or ahistorical explanatory functionality. The current utility specifies what the behaviour does from the perspective of the entire group. A behaviour may have a current use function for the group even if this function has no historical role in evolution or significance to understanding how the group works, but it is the starting point of analysis. The historical evolutionary functional explanation would make a stronger claim that the behaviour exists because it has evolved for its group-level function. The ahistorical evolutionary explanation makes the claim that the behaviour has a causal function as a part of a complex system that can be understood on the group level only, as a survival strategy of the entire group.

There is a degree of such collectivism in human groups thanks to culture. Biologist Mark Pagel (2012) has argued for a superorganismal view of human cultural groups. According to him, cultures and

especially languages bind the individuals of the group together and differentiate groups from each other in ways that makes them effectually superindividuals. This would be an extreme case of collectivist holism in the proximate dimension and would make cultural groups individuals and also the central units of selection. Others who treat cultural groups as possible units for selection usually consider some particular traits to be group-level adaptations and some others (most) individual-level adaptations (see for example Boyd & Richerson 1985 & 2005; Wilson 2002; Sterelny 2012). I will take the latter approach to be more realistic. Something being cultural does not automatically make it a group-level property either, but culture is what can make human groups collectivist if anything does. Culture-specific norms, normative expectations, and meanings may cause behavioural uniformity or orchestrated division of labour. The main significance of culture as a uniformizing force is not in the proximate dimension, however, but in the developmental dimension. Even if collective or interactive traits exist because of cultural factors, these factors are not independently existing proximate causal factors, but instantiated by individuals. Cultures influence behaviour by influencing what behavioural traits (including reactive attitudes towards the other members of the group) individuals acquire. It is time to move to the developmental dimension now.

7. Evolution, Replication, and Development

The developmental dimension of evolutionary explanations is an important topic for the issue of evolutionary individualism and holism in two different ways. First, the individuation of the replication process tells us *what* is being selected. Second, there is a direct connection to the proximate dimension. The trait must be replicated to evolve, and the details of its developmental processes and its interaction with the environment are important in understanding how this happens. Some of the mechanisms of inheritance are "external" to the developing organism and this opens up the possibility of supra-individual developmental mechanisms. Human culture is a candidate for such a mechanism. Culture introduces the social aspect to all stages of development, from the development of psychological capacities to learning and behaviour modification as an adult. This can make the spreading of traits horizontal within the group of interacting individuals instead of vertical (in parental lineages only). This, in turn, can make the evolution of social behaviour holistic in the sense that the competition between traits does not depend on only individual fitness effects but also on what traits spread socially. At the same time, the groups that are bound together by having acquired the same trait share the fitness consequences. This is the cultural group selectionist link between culture and fitness (Boyd & Richerson 1985 & 2005). The cultural origins, evolution, and maintenance of a behavioural trait also facilitate the interactionist traits I discussed in the previous chapter. As I concluded there, culture might even unify groups in a way that makes them collectivist in all three dimensions.

The role of culture in evolution is, however, much more complicated than the "memetic view" in which culture is simply a different (possibly competing) medium of inheritance to the genetic inheritance. The received view of the logic of evolution has genes at its centre as *replicators*, and some views of cultural evolution build upon this framework. This received view has been challenged over the past two decades, mostly because of developmental considerations. The

developmental challenges have created a new field of evolutionary developmental biology ("evo-devo") that has become an established field bridging evolutionary and developmental biology over the past few decades. The developmental challenge has been accompanied by new discoveries about the evolutionary process itself. Some of these challenges, discoveries and theoretical advances have been combined under the umbrella of the *Extended Synthesis* by some, calling for a paradigm shift in biology. Although the new framing has remained a minority position within evolutionary biology, its central insights are generally accepted as expansions to the Modern Synthesis at minimum. Following many philosophers of biology, I consider the new paradigm to be worth taking seriously, especially in the human context. I will also discuss the Developmental Systems Theory and its reconceptualization of evolution, based on developmental considerations. The latter is a much older and much more radical suggestion for a paradigm shift, but its central insights are partly similar and, to the extent of detail relevant to my main topic, combinable.

The Extended Synthesis has consequences for how we understand the role of social interaction and culture in evolution – and what culture is. A central topic in the new paradigm is the extended perspective on what factors participate in producing phenotype and what subset of them can function as a route for inheritance. Selection can work on some factors in the development, but it cannot work on others. What is important is the causal connection between a transferable factor and a specific difference-making outcome in the developmental process. Some of these factors can be shared by multiple individuals in a connected way that binds the replication of the traits that these factors contribute to produce. I call this "developmental holism" since the factors that are of interest here are part of the individual developmental process, but they are factors that bind individuals together. This contrasts with developmental processes where all the developmental factors that are also routes of inheritance are *individualistic* routes of inheritance. These traits are *innate*. This is a contested concept and a contested idea, but I will argue for it and show that the core idea of nativism in nativist

evolutionary psychology is sensible – although it does not have its presumed consequences. I will do that in the next chapter.

The emerging view about the connections between evolution and development also changes the biological understanding of culture and its role in evolution. Culture is an ambiguous concept, especially if it is given an explanatory role; I will make a few remarks about this later. But all things cultural are part of the non-genetic inheritance. I will next discuss the idea of the replicator and why genes are not it; the Developmental Systems Theory as a challenger to the gene-centred view; and the paradigm shift to the Extended Synthesis, and how the other forms of inheritance figure in, after which I return to culture specifically. The existence of these mechanisms alone does not imply the horizontal transmission of traits, which would make the spread of traits a holistic process, as cultures are supposed to be. But it might. Furthermore, some of the social interactions that affect development are about the modification of capacities, attitudes, behavioural habits, skills, and so on, in ways and at developmental stages where they cannot be conceptualized as transmission of memes, representations, or any other mentalistic entities. From the biological point of view, culture is not one thing or one kind of mechanism, but a cluster of things and developmental mechanisms.[199] Culture has a capacity to make cultural groups holistic, but the horizontality of transmission is a separate issue from external inheritance.

7.1. Taking Development Seriously

The classic *Modern Synthesis*, so termed by Julian Huxley (1942) and referring to the synthesis of Darwinian evolutionary theory and Mendelian genetics, two fields that were long considered rivals in the study of population-level change (see Depew & Weber 1996), left

[199] There are, of course, attempts to define culture nevertheless. I will come back to this.

developmental processes in a black box. The basic idea was that it is enough to know of the existence of inheritable factors that participate in both developmental and evolutionary processes, and the systematic connection between the transferable entities and the phenotypic traits is enough. The gene-centred view essential to classic models of social evolution is based on the replicator-centred view on the general logic of evolution as explicated by Richard Dawkins (1976) and David Hull (1981 & 1988a). The developmental details matter, however. The epigenesis is an orchestrated process where the existing stage of development determines which genes and environmental factors affect the next step, and a multitude of these factors determine what the next step is. Furthermore, the genetic "instructions" are "about" the developmental steps, not about the resulting phenotypes. (Wolpert *et a*l 2010.) It therefore seems that genes do not fit the ontological role given to them as the basic unit of copying itself and being the ultimate unit of the evolutionary process (as Dawkins 1976 and other genetic selectionists would have it; see also Sterelny & Kitcher 1988).

Accordingly, the gene-centred Modern Synthesis has been challenged over the past two decades – for many reasons, but mostly because of developmental considerations. The genes do not fulfil the role they were supposed to play (see Oyama 1985; Oyama, Griffiths & Gray 2001; Moss 2004), and the environment is an active part of the evolutionary process (*ibid;* Odling-Smee, Feldman & Laland 2003; West-Eberhard 2003; Jablonka & Lamb 2005). This has created new interpretations of the evolutionary process that push the genes from centre stage. These approaches include the Developmental Systems Theory (Oyama 1985; Griffiths & Gray 1994; Oyama, Griffiths & Gray 2001; Gottlieb 2001 & 2003), Peter Godfrey-Smith's *Neoclassical Individualism* (2009), and the Extended Synthesis. The latter brings together various new insights into the evolutionary process, such as *niche construction* (Odling-Smee, Feldman & Laland 2003), evolutionary developmental biology (Hall 2003; Carroll 2005; Müller 2007), and alternative routes of inheritance, labelled "soft inheritance" by Ernst Mayr (1980; see Jablonka & Lamb 2005 & 2008), reframing the general view

of how evolution works. The difference between replicators and interactors is central to the logic of evolution, however, and to conceptualizing the levels and units of selection. I will return to this shortly; now I will take a closer look at some aspects of evolution and development that are relevant to making this point.

7.1.1. Replicators, Interactors, and Developmental Systems

The debate over *gene-selectionism* (the gene as the basic unit of evolution; Dawkins 1976, 1984; see also Sterelny & Kitcher 1988) and the debate over the levels of selection, as in whether only individuals can be the loci for selective forces or other levels of biological organization can as well, are two separate issues (see Sober 1984; Sterelny & Kitcher 1988; Lloyd 1992 & 2001). Robert Brandon (1982 & 1988) distinguishes between *units* and *levels* of selection. The unit is an entity that is selected, and the level is the level of organization that has the properties relevant to the selection. I have already mentioned that I will make a further distinction between the level of selection and level of adaptation. The last chapter argued for holistic adaptations in the form of interactionist traits while trying to be neutral regarding the levels of selection; so far, the argument regarding them has only been that the *traits* selected are created in the interaction between individuals, but this leaves open the possibility that the relevant selection processes work through individual fitness only. This is a separate issue. The issue of units of selection, however, is how we should think about the entities that are replicated: genes or individuals.

In the gene-centred view, Brandon's distinction between units and levels coincides with Richard Dawkins's distinction between replicators and vehicles, refined by David Hull, whose terminology of replicators and interactors is more popular among philosophers (Dawkins 1976; Hull 1981 & 1988a). Replicators are entities that are replicated, and evolution happens because of their fitness differences, while interactors are the entities through which the replicators interact

with the environment and cause the fitness differences. In an individualist framework, an individual is both a replicator and an interactor, but the roles are separated in the logic of selection. One way to conceptualize group selection is that groups act as higher-level interactors (Brandon 1988; Sober & Wilson 1994). In principle, any evolutionary process can be described both as the evolution of interactors and as the evolution of replicators (Sterelny & Kithcer 1988), and if gene-selectionism were true, the gene's-eye-of-view of social behaviour would be easily combined with a multilevel selection model (see Wilson & Sober 1994; Okasha 2006; Ågren 2016). Both issues (the identity of replicators and the nature of interactors) are contested.

Our knowledge of genes and individual development has increased enormously since the debates over replicators and interactors first emerged. The fundamental problem with the simplified image of genes is that the molecular genes that participate in the developmental processes and the "genes" of population genetics and evolutionary theory are different both conceptually and substantially (see Moss 2001 & 2004; Stotz & Griffiths 2004; Griffiths & Stotz 2008 & 2013). Molecular genes, or Lenny Moss's "D-genes" (Moss 2004), are defined by their causal powers in the developmental process and the population-level genes, Moss's "P-genes", are defined by a phenotypic product that is genetically transferred to the next generation. D-genes refer to a type of cause (with whatever effects these causes have) and P-genes to whatever causes certain effects. There is no one-to-one causal relationship between D-genes and P-genes. The phenotypic properties are produced through combinations of D-genes, with different combinations possible for any given property. A D-gene "for" a property can only be an abstracted invariant dependency between the molecular gene and a property when other genes and environmental factors are kept fixed in an explanatorily sensible way. An explanatorily sensible way might be, for example, to fix all the genes that are shared by the whole population and are therefore of no practical consequence. Another sensible omission would be to concentrate on only those genes that have a specific effect on the phenotype (the **instructive**

factors) and omit the genes that are only necessary components for the development to take place (the **enabling** factors; see Gilbert 2003; Griffiths & Stotz 2013; Calcott 2017).

Nevertheless, a molecular gene "for" a property is an abstraction that only makes sense in a context not only of other genes (that participate in the causal process) but also of equally necessary environmental factors, even if they are only enabling. Another problem is that there are also instructive environmental factors. For example, the sex of some reptiles (crocodiles and turtles, for example) is determined by the temperature of the sand where the eggs are buried. This all becomes much more relevant when we address behaviour, especially human behaviour. Moreover, if the D-gene "for" a phenotypic effect is whatever is the difference maker, it may actually be the *lack* of a molecular gene. In addition, P-genes are discovered as heritability patterns on population level. They are basically theoretical constructs for studying heritability on the population level – they cannot be the ontological basis of evolution.[200]

Epigenesis is a holistic process in which genes and environmental factors interact. The various factors are not additive but interdependent in their functions, and the interactions are often non-linear and cyclical. The genes and other factors are causally symmetrical in this process.[201] From the developmental point of view, genes are important

[200] The standard defence of gene-selectionism was that precisely because evolution works on the population level and genes are replicators, only correlation between the genes and all phenotypic traits that it is associated with on the population level matter. (See Sterelny & Kitcher 1988.) This, however, does not work for exactly the reasons why we need to distinguish the two notions of gene.

[201] Even if both genes and environmental factors affect height, we cannot say how much of the height of a given person is caused by genes, nor how much by environmental factors. The causal mechanisms determining the process on the individual level are too complicated. Relative measures can be used on the population level, but population-level measures do not measure the developmental processes. The heredity of a trait basically measures the relevant genetic variation relative to the relevant environmental variation as they exist in

but not unique factors. From the evolutionary point of view, what makes genes important is that they are a powerful way to transmit phenotypic properties relatively reliably. Again, they are important but not unique; non-genetic factors may also serve as developmental resources that can transmit phenotypic properties (see Griffiths & Gray 1994 & 1997; Sterelny 1996; Oyama *et al* 2001; Odling-Smee *et al* 2003; Jablonka & Lamb 2005; Shea 2011; Tikhodeyev 2018). This is why individualist models of replication have emerged: it is the individual itself that is a replicator. One such model is Peter Godfrey-Smith's generalized Darwinian model of evolution, which is based on *Darwinian populations* that consist of *Darwinian individuals* that are capable of Darwinian processes (Godfrey-Smith 2009). These individuals do not necessarily coincide with individuals as *organisms* defined by metabolism (Dupré & O'Malley 2009; Godfrey-Smith 2013); instead, they are reproducing continuants. Godfrey-Smith's theory abandons the gene-centred view of evolution and is critical towards it, but it is a formal generalization of Darwinian logic and does not engage in developmental processes as such. The sophisticated individualist theory based on development that we must look at more closely in this context is the Developmental Systems Theory.

The originator of both the term "developmental system" and one of the foundational ideas of the Developmental Systems Theory, development as a system of interactions with emergent properties that cannot be reduced to individual genes, is Conrad H. Waddington (1952). The theory as we know it now has been most prominently developed by Gilbert Gottlieb (2001 & 2003), Susan Oyama, Paul Griffiths, and Russell Gray (Oyama 1985; Griffiths & Gray 1994, 1997 & 2005; Oyama *et al* 2001). According to the theory, the units of evolution are not genes but not phenotypes either, but whole *life cycles* of individuals, the developmental systems. A developmental system does

---

the population at the moment of the measurement. What genetic and environmental factors are relevant and to what extent depends on the developmental processes, and the population-level studies are silent about these.

not, however, replicate itself, but transfers developmental resources that (with other reliably present developmental resources, including those in the environment) *reconstruct* a similar next-generation developmental system. The resources include genes, organelles (such as mitochondria), the system for gene expression (the mother's sex-cell from which the new individual starts growing), but also all the environmental factors that explain why the next generation resembles the previous. This may include the social and cultural environment. It is an almost trivial fact that external factors play a causal role in development, but what the Developmental Systems Theory emphasizes is that non-genetic factors are not simply participatory or *enabling* factors, but sometimes *instructive* factors that have a specific effect on the phenotype and therefore participate in the construction of the phenotype.[202] This construction process is holistic and has a high degree of contingency in it. Furthermore, the developmental system does not have an end-state or a "result phenotype"; rather the developmental process continues throughout the life cycle. This gives a natural evolutionary perspective on the interaction between plastic phenotypes and the environments in which they produce adaptive traits. The traits whose development is influenced by or even depend on certain specific features of the environment may be adaptive in the ecological environments where these features exist (see Gilbert 2001 & 2003).

Obvious problems with such a view are its holism and the fuzzy borders of the developmental system (Sterelny 1996), running the risk of "dissolving into an interactionist causal soup" (Shea 2011). The distinction between instructive and enabling factors aims to bring clarity to this. Another solution is to approach development as a process that

[202] This could be conceptualized as non-genetic factors transferring information or "inherited representations" (Shea 2011), but the notion of biological information raises a whole range of problems and is rejected by most Developmental Systems theorists as highly misleading even (or especially) in the case of genes (for example, Oyama 1985; Griffiths 2001; Moss 2001). The notion of causal specificity discussed in chapter 2 is probably sufficient (see Calcott 2017).

goes through a mechanistic structure where the parts of the mechanism are the developmental resources, which enables us to give *qualitative* information about the causal structures and specific nature of relevant factors in the development (see also Griffiths & Tabery 2013). Developmental biology traces developmental processes and how the various factors (such as genes and environment-induced hormonal changes) interact within epigenetic processes that have a certain mechanistic structure, and the field can be understood as being about discovering mechanisms.[203] The mechanisms are self-building – earlier steps of the development enable later ones, not only because the process progresses in stages, but because the earlier stages build mechanisms that can produce later stages. William Wimsatt (1986a, 2001 & 2007) calls this layered productivity *generative entrenchment*. I will return to the virtues of the mechanistic approach later. Psychological development is an expansion of this process (see Griffiths & Gray 2005).

I will not discuss the Developmental Systems Theory in detail, but I assume that something along this line is the correct way to understand the replication process: individual genes are not an adequate focus for understanding replication from the evolutionary point of view. The individuals – who are produced by different developmental resources and pass them on – are. This seems to be the general consequence of what we know about how biological development is relevant to evolution. There may be other ways to conceptualize this, but the Developmental Systems Theory is the most sophisticated and

[203] Developmental biology does not have a field-defining theory, but it uses a cluster of theories, models, and other tools, some of which have originated within the field, some of which have come from elsewhere. Alan C. Love (2014) has proposed that to understand the structure of the field, the perspective should not be from theories but from the problems that organize the field. This fits nicely with the view of scientific inquiry proposed by Matti Sintonen (1984, 1989 & 1990) and Jaakko Hintikka (1988). A compatible way to understand developmental biology that I suggest here is that the development of an organism is a mechanistic process, and the various research questions that arise are about different components of the mechanism.

integrative model available. This position is also more problematic for group selection than the mainstream gene-centred view of social evolution, which means that the case for the significance of the group-level considerations becomes stronger if it is shown to be needed within this framework too. In the gene-centred framework, the fitness of a gene can be correlated with phenotypic differences on any level of biological organization, at least in principle,[204] but the individualist framework requires more work. In most cases, the individual is also the interactor, seemingly breaking the distinction between replicators and interactors. But the logic that the distinction articulates does not break. There is still interaction between the individuals and environments and there are replication processes in forms of transfer, although some forms might take external routes that are also forms of interaction, such as modifying the environment; I will return to this shortly. More importantly, even if replicators are individuals, interactors may still be realized on several levels – or rather, the interactions with the environment that determine reproductive success may take place on the level of interactions between individuals or in their collective behaviour. As I emphasized in the previous chapter, the interactive traits are instantiated by individuals and their interaction, and although the traits in the selective competition are holistic, they can be replicated through individuals, and the issue about the levels of selection, which is the third dimension, is left open by these considerations only. Before going into individualism and holism in development more closely, I will now briefly review some of the more precise connections between evolution and development in greater detail.

[204] This can be derived from Price's Equation (Price 1970). Covariance it shows is not, however, sufficient to characterize group selection, for it cannot handle soft selection (Wallace 1968) where each group determines the number of offspring to be the same, which seems to be a case of group selection but without fitness variations caused by groups (Okasha 2006). However, it shows how group-level selection is possible in the gene-centred view.

7.1.2. Evolution and Development

Psychological and behavioural traits do not fall into fixed (innate, genetically determined, or what have you) traits and completely acquired traits that are imprinted without predisposition; a lot of development possesses a plasticity in which both "intrinsic" and environmental characteristics affect the development in ways that are evolutionarily functional (see Gilbert, Opitz & Raff 1996; Gilbert 2001, 2003 & 2007; West-Eberhardt 2009; Pigliucci & Müller 2010; Wolpert *et al* 2010). Selection is not only selection of phenotypes, but also the developmental pathways that lead there. Selection guides the pathways to rely on the developmental resources that are reliably present when fixity is needed – if environmental resources are reliably present, a pathway using them is enough and there will be no selection for dependence on genes or selection for any particular genes associated for this. The emergence of reliably present, adequate environmental factors can even break genetic control over a developmental process. On the other hand, the emergence of an adaptive phenotype through phenotypic plasticity, including learning, can lead to genetic assimilation of the trait (*phenotype first* evolution; see Waddington 1953; Piggliucci *et al* 2006) and the fitness advantages from fast learning increases selection pressure on genetic evolution of faster learning and growing genetic control of the development (the *Baldwin Effect*; Baldwin 1896). However, if the environment contains variation that makes different phenotypes more adaptive in different contexts, there may be selection for *phenotypic plasticity*.

There are two senses of phenotypic plasticity. It can be the organism's ability to react to an environmental input (internal or external) with a change in form, state, movement, or rate of activity. That is, the organism can change or behave differently depending on the context. It can also mean the ability of a single genotype to produce more than one phenotype. That is, the environmental factors can trigger different developmental pathways such that a factor that is robustly connected to an adaptive need triggers a pathway that leads to a phenotype that

is adaptive in this environment. The trigger does not need to be informatively connected to the challenge or the adaptive form – selection operates with robust causal connections, not connections in the content. This is true of both behaviour guidance and development. In complex and hostile environments, it becomes more useful for behavioural guidance to become more precise with the connection between the relevant features of the environment and the perceptual data and to use multiple cues, decoupling the representation of an external thing from direct perception and moving gradually from stimulus-response to cognition and robust motivational mechanisms (see Godfrey-Smith 1996; Sterelny 2001a & 2003). A similar process takes place in the plasticity of development: some features of psychological development become more sensitive to more precise informational content, and as the behavioural responses become more flexible and perhaps also include imaginative solutions, we can talk about learning. Regardless of whether general, universal learning exists or not, there is a gradual shift from causally triggered plasticity and powerful flexible learning, and most human learning is biased and constrained, even symbolic learning (Sperber 1996; Deacon 1997).

From the evolutionary point of view, developmental processes that are fixed or have robust, predisposed developmental pathways that have specific triggers are more reliable in producing adaptive behaviour but require relatively stable environments and enough time to evolve. Individual learning is faster and more suitable for changing environments but is constrained by individual epistemic and cognitive limitations, and the cumulation of feedback from the environment and success is limited to one lifetime. (Sterelny 2003; Ylikoski & Kokkonen 2009.) Social and symbolic learning lie somewhere in between: they are faster than genetic evolution but slower than individual learning, and they allow multi-generational cumulation but make individuals vulnerable to mistakes by the individuals they copy from. All three ways to adapt are good in some cases – cultural adaptation very rarely. (See Boyd & Richerson 2005; Richerson & Boyd 2005; Sterelny 2006a.)

The overall point here is that the interaction of the internal and external factors in a specific way may be selected. Adaptations are not only passive adaptations of phenotypes to environments, but adaptations to utilize environmental features, making the products of these interactions extensions of phenotypes (as suggested by Dawkins 1982), but also in development. When organisms modify their environments, they modify both the developmental environment for future generations and their selection environment, causing loops and co-evolution between the environment and the population inhabiting it. John Odling-Smee and his collaborators dubbed this process *niche construction* (Odling-Smee *et al* 2003). Some aspects of culture can also be understood as niche construction (Laland *et al* 2001; Odling-Smee *et al* 2011). Humans in particular change their developmental conditions actively, including social and cultural contexts, the skillsets that a child needs to learn, and so on. Human behaviour is both very flexible and suggestible. This brings to the fore the importance of understanding social and cultural aspects of development not only in evolutionary anthropology, but also in evolutionary psychology.

At least some human psychological characteristics have a developmental connection with other people, and their evolutionary function is related to the nature of this connection. I discussed nativist evolutionary psychology in a previous chapter and noted that there are alternative approaches. The styles of evolutionary psychology are related to one's view on how evolutionary theory works. Karola Stotz, for example, has argued that nativist evolutionary psychology is a natural application of the (outdated) Modern Synthesis and has suggested and outlined an alternative based on the Extended Synthesis and the Developmental Systems Theory instead (Stotz 2014).[205] Darcia

[205] Furthermore, nativist evolutionary psychology assumes the "old school" of cognitive science, while Stotz suggests that the 4E movement (embodied, embedded, enactive and extended cognition) is a more suitable framework for understanding the object of evolutionary psychology. Generally speaking, evolutionary psychologists and Developmental System theorists have been antagonistic towards each other, although attempts other than Stotz's to

Narvaez, with her collaborators, has in turn studied child development and how social and other environmental factors and parental styles affect the development of empathy, intelligence, social emotions, and psychological well-being, using evolutionary hypotheses about the evolutionary contexts of evolved psychological development as a theoretical framework (Narvaez 2014; Narvaez *et al* 2012, 2014 & 2016).[206] Evolutionary psychology does not need to be methodologically individualist in its approach to cognitive processes, in its assumptions of psychological development, nor in its views about evolutionary processes. However, acceptance of holism does not always entail the futility of individualist approaches in any of these dimensions. This is a case-by-case issue. I will return to this in the next chapter.

7.2. Individualism, Holism, and the Extended Synthesis

The Extended Evolutionary Synthesis is an attempt to produce a new synthesis of evolutionary biology and other fields of biology that are relevant to understanding evolution, especially developmental biology, but it also seeks to re-think evolution in the light of more recent discoveries in evolutionary biology proper. It builds heavily on evolutionary developmental biology (see for example Gilbert 2001 & 2003;

---

integrate the two projects exist (for example, Frankenhuis, Panchanathan & Barrett 2013).

[206] Narvaez integrates results from developmental and child psychology, anthropological research of hunter-gatherer childhood, and comparative studies of how other social mammals raise their offspring. According to her, there is an evolved developmental niche that she calls the *Evolved Nest*. For babies, it includes immediate responsivity to crying (no distress to babies), breastfeeding for 2-5 years, and frequent touch; later in early childhood, it includes free play in natural world with multi-aged playmates, multiple adult caregivers, and a positive social climate. Deprivation of these features is deprivation of essential developmental resources.

Hall 2003; Carroll 2005; Müller 2007), which integrates developmental biology into the Modern Synthesis. In addition to the connections discussed above, it seeks to understand developmental processes from the evolutionary point of view, and things such as developmental constraints and homologies, by integrating the two fields. The Extended Evolutionary Synthesis aims at a much more fundamental paradigm change. It does not suggest a radically new ontology for evolution like the Developmental Systems Theory does, but other than that, these two approaches are compatible at the level of detail needed here. I will now provide a brief summary of the Extended Synthesis to give a systematic image of how developmental mechanisms interact with evolutionary mechanisms, after which I will rephrase the individualism–holism issue in developmental dimension within this context.

7.2.1. The Extended Synthesis

The Extended Evolutionary Synthesis (Piggliucci 2007; Müller 2007 & 2017; Piggliucci & Müller 2010; Laland *et al* 2015) seeks to bring the core logic of the evolutionary theory up to date with recent discoveries and theoretical advances in biology across its subfields. Many of these ideas are present in "mainstream" evolutionary biology, and the critics of the paradigm shift idea do not so much refute any component of the Extended Synthesis as they downplay them as "add-ons" to the Modern Synthesis. They remain agnostic on how important these might be rather than either denying them or wanting to change the paradigm (see Wray, Hoekstra and colleagues in Laland *et al* 2014; see also Müller 2017). I take the Extended Synthesis, therefore, as being relatively non-controversial in its substance and important in the human context especially, given that all its "add-ons" are important in the human context. Furthermore, the "add-on" approach is telling of the theoretical status of evolutionary biology today: as Gerd Müller (2017) points out, it is a set of various topics, approaches and research programmes, not a theoretically unified field. We need to see how they all fit together – and what consequences this has.

The elements that the Extended Synthesis aims at integrating include the direct evolutionary significance of non-genetic or "soft" routes for inheritance (Mayr 1980; Jablonka & Lamb 2008) such as epigenetic (see Jablonka & Lamb 1995 & 2005; Tikhodeyev 2018), behavioural (including nurture, social learning, and transforming the environment; see Avital & Jablonka 2000; Jablonka & Lamb 2005; Hoppitt & Laland 2013), and symbolic (Jablonka & Lamb 2005); the understanding of developmental processes as non-linear processes that both constrain and facilitate selectable variation (West-Eberhardt 2003; Gilbert 2007; Wolpert *et al* 2010); the forms of phenotypic plasticity, including phenotype first sources for adaptation such as the Baldwin effect, the consolidation of epigenetic changes, or genetic assimilation (Jablonka & Lamb 1995; West-Eberhardt 2003; Piggliucci *et al* 2006); and widening evo-devo into *eco*-evo-devo with niche construction (Odling-Smee, Feldman & Laland 2003; see also Gilbert 2003). The proponents of the Extended Synthesis are less biased towards natural selection explanations when it comes to evolutionary explanations proper, but within natural selection explanations, they tend to accept multi-level selection (Müller 2017).

According to one systematization of the Extended Synthesis (Laland *et al* 2015), evolution proceeds in the following way. There is a population of developing organisms (or developmental systems). The developmental processes are influenced by genetic, epigenetic, and environmental factors. The relationship between genes and the development is not one-directional: the developmental processes affect the gene expression. The factors that contribute to inheritance include genes, epigenetic mechanisms, modification of ecological factors, and culture. Genetic processes such as mutation and recombination are only one source of evolutionary novelty. For example, epigenetic processes can affect gene expression in a novel way. *Environmental induction*, where new environmental factors create new phenotypes within the existing developmental possibilities due to developmental plasticity, can lead to new variation and potentially adaptation through

selection.[207] The behaviour of organisms can create new environmental conditions (niche construction). Finally, in *phenotypic accommodation*, developmental plasticity leads to novel adaptation through adaptive adjustment, where the negative effects of a mutation or environmental induction are balanced by developmental processes (see West-Eberhardt 2005; Badyaev 2009).[208] The processes that modify the frequencies of heritable variation (genes and other) include selection (working on multiple levels), drift, and gene flow, but there are processes that are not a part of the Modern Synthesis which bias the selection. One is *developmental bias,* which includes developmental constraints but also *developmental drives* that push the development towards some directions (instead of preventing some directions), which might even facilitate adaptive evolution since the selection may shape the developmental bias itself (Uller *et al* 2018). Another biasing factor is niche construction, in which the organisms modify the selective environment, not only the other way around. These things together determine *phenotypic evolution*, the change in the phenotypes and their frequencies on the population level.

Now, the Extended Synthesis has four general consequences that change the image of evolution from the Modern Synthesis. First, it completely changes the role of genes, just like the Developmental Systems Theory. Neither denies their importance, but both deny their uniqueness as a developmental resource or as a route of replication. Consequently, the definition of evolution as a change in the gene pool of the population does not work – and it would be misleading to understand evolution as a change in the living world understood in

[207] This can be the case with some invasive species, for example, and it can lead to significant fitness advantages, making it a starting point for adaptation. See Badyaev & Oh 2008.

[208] Phenotypic accommodation is not caused by the novel factors, it is not random, and it is not systemic self-organization. It is an evolved property of the developmental system. The accommodation is caused by ancestral adaptive responses, making the new phenotype viable, but the phenotype itself may be new. (West-Eberhardt 2005.)

terms of phenotypes, since phenotypes can change without genetic change *and* genes can change without phenotypic change. Second, the details of the developmental mechanisms matter with regard to the units of evolutionary change within an organism. These mechanisms are, however, evolved properties, whether selected or not. I have referred to some of the results of evolutionary developmental biology such as environmental induction and developmental drive, where the effects of developmental systems on further evolution are because of adaptive selection of developmental processes in the past. The interplay between evolution and development is bidirectional and even the Developmental System theorists give selection a central role in shaping the developmental architecture of the developmental systems (see Griffiths & Gray 2005), although they emphasize contingencies and the fact that the new developmental system is *constructed* from similar resources every time – there is no predetermined, inherited *plan* for how this takes place. A special case of such processes that will soon be relevant to the discussion is *canalization* (Waddington 1942 & 1957). Some phenotypic properties are such that they emerge in various combinations of developmental resources. If a factor that is usually present and active in the development is missing, another developmental pathway emerges that results in the same phenotype. This is selected robustness of the outcome of crucial characteristics. Third, the environment plays an active role in producing the phenotype. I stated before that selection is not selection of phenotypic traits alone, but also the developmental mechanisms that lead to those traits (see Gilbert 2001 & 2003). Environment is a part of this and should be included in the interactions that are selected, especially if there is looping in the form of niche construction. Fourth, the Extended Synthesis is an attack on adaptationism (see Müller 2017). Again, natural selection is important, but it is not the only source of adaptivity. Not direct, at least.

I have already agreed about the genes above, and I elided the role of genes in the previous chapters when I was discussing evolutionary explanation and defining altruism. Heritability is essential to

evolution, but its route is not, which makes genes as such irrelevant to the general logic of evolution. Genes are still the most important route of inheritance, but the changed roles of the gene have some consequences. First, the theory of kin selection requires re-interpretation. Second, if there are different systems of inheritance, it is theoretically possible that there are different fitnesses: the capacity to replicate takes different routes and different routes may have different competitions between individuals. In practice this is almost never the case if individual is the replicator, and all the different inheritance systems participate in the development of this one entity. Culture, however, is an exception to this, potentially making human evolution different. How some culturally transmitted components of psychological development spread is detached from other dimensions of inheritance.

I also *disagreed* about adaptationism before. Recall that I distinguished between historical and ahistorical explanatory adaptationism. The developmental phenomena discussed in this chapter makes the assumption of natural selection having been the primary causal factor in the evolutionary history of an adaptive trait problematic, among other previously mentioned reasons, and I alluded to these developmental factors earlier as well. However, the argument I made about understanding organisms as functional wholes in their environments from the point of view of adaptivity still holds. First, the origin and some evolution of adaptive traits because of developmental phenomena is a competing historical explanation, not a competing view of adaptive functionality. Second, selection participates even in these processes, as explicated above. Adaptively dysfunctional characteristics are still selected against, and functional characteristics are selected for – it is just that the entire process is more complicated than in the Modern Synthesis. Third, many of the more complicated developmental mechanisms that are significant for evolution are evolved mechanisms themselves. Generally speaking, natural selection plays enough of a role in the process to make adaptivity an adequate perspective on the functionality of organisms and their behaviour. The other two consequences, the centrality of developmental mechanisms and the active

role played by the environment, are central to the main topic. I will now define developmental individualism and holism, building on these considerations.

7.2.2. Developmental Individualism and Holism

The development of an individual organism is an interactive process between various factors that takes place in stages. Some of the stages are there only to enable later stages. William Wimsatt's concept of *generative entrenchment* refers to the phenomenon in which earlier stages of the development are generative for the later ones but are entrenched in the process, so that if these processes get disturbed, the later stages are disturbed too, potentially making the organism inviable. The more deeply entrenched the developmental process is, the more likely it is to be canalized (as an evolutionary response to vulnerability), making it more resistant to changes in both genome and environment. This resistance to change, together with the fact that change would make the organism inviable, makes these developmental processes and the phenotypic characteristics they produce more constrained with regard to evolutionary change. (Wimsatt 1986a, 2001 & 2007.) As discussed above, different factors that play a *causal role* in the development play different *kinds* of role in terms of what exactly they affect in the phenotype: some are mere enabling factors and some are instructive.[209] Within the enabling factors, there may be factors necessary for the development and factors that are alternatives to each other because of canalization. Among the instructive factors (that cause specific changes in the phenotype), there are differences in the scope of change in the characteristic affected. We can conceptualize the developmental process as a *self-constructing mechanism* (or sets of

[209] There is no reason that "instructive" and "enabling" should be clear-cut categories. There is, however, selection for both sensitivity and insensitivity to developmental factors regarding their effects. Both robustness and plasticity can be selected and the very systems of inheritance are evolved systems.

mechanisms) and the various factors as components of the mechanism, at different stages of the process. This process is holistic, with complicated interdependencies, and we cannot quantify the relative importance of the various components, but we may nevertheless discover the qualitative structures of the causal mechanisms that connect the factors to the development.

Not all parts of the developmental mechanism are equally important to any given explanatory purposes. This depends on three things. First, the facts about the causal system. For example, instructive causes are more important to some questions than enabling causes. Second, they depend on the explanatory interest. Developmental biology is (and should be) interesting in all factors, but not everything is necessary to evolutionary biology or to psychology. Third, not all the range of variation in the environment is important. For example, the factors that are necessary for the development to take place at all (such as the right temperature or atmospheric conditions) are not relevant for most considerations. The range of variation depends on the explanatory questions. It may be, for example, theoretically interesting to know how a plant grows in zero-gravity, but for most purposes, it is reasonable to limit the relevant variation to actual or plausible ecological conditions. If we are interested in the relevance of developmental processes to the evolution of a particular trait, the factors that are always a part of the developmental environment can be blackboxed. They do not generate phenotypic variation even if they are a necessary part of the development. Furthermore, we can stipulate normal conditions even if the conditions do not always hold. There can be change events, for example, that disturb development but are rare and should not be considered as part of the normal range of the developmental environment. For example, it is possible that a child grows up without any linguistic environment at all, and this has happened and been recorded several times, but any theory of language acquisition works under an assumption that a linguistic environment exists. Sometimes it is more sensible to treat an omission as an explanatory cause. For example, parental care matters both in animals

(Klopfer 2001) and humans (Narvaez *et al* 2012; Berk 2013) and understanding exactly how is important, but for some explanatory purposes, *neglect* of something that is normal is the change that explains *abnormality* in development. I will return to these topics in the next chapter where I discuss the concept of innateness and what theoretical or explanatory role considering something innate could have.

From the evolutionary point of view, those parts of the developmental mechanism that are inherited play a special role. Those environmental factors that are present in all relevant environments are causal background only, even if they are part of how the phenotype is constructed. The same applies to the parts of the environment that vary and affect the phenotype through phenotypic plasticity but are not transmitted to the next generation. Take the sex determination of some reptiles by temperature for example. The development of both sexes are evolved processes, and which temperature is the borderline between different triggers is probably an adaptation for ensuring the optimal frequency of different sexes in the given environmental conditions. The mechanism that produces two sexes is selected as a whole, and the varying environment is a decisive part of it, but only the parts of the mechanism that depend on the parents' contribution to the developmental process are important from this perspective. This may sound at odds with Developmental System Theory, which holds that every generation of organisms is reconstructed and all the developmental resources are equally important, and the environmental factors that construct the next generation are a part of the inherited developmental resources that evolution builds the developmental systems on. I do not deny any of this. What I am saying is that all the parts of the mechanistic process that constructs the organism are important for understanding *evolution*, but hereditary resources are the only parts of the replication process that are important for *selection*. This difference is important.

Consider a highly general and flexible learning process as an example of one end of plasticity. If this capacity (C) is an adaptation, it has evolved for an environmental variation ($E_{1...n}$) regarding the

context of learning. There is a multitude of possible learning results ($L_{1...n}$). C is selected on the condition that the choice of $L_{1...n}$ in the given E is more likely to be more adapted in the E than most of the other $L_{1...n}$ or any fixed variant. This adaptive matching between E and L is the adaptive consequence that the learning is selected *for*. However, the capacity itself and whatever processes pass this capacity on to the next generation are what are selected, not the set of resulting behaviours $L_{1...n}$. No phenotype is without *some* of the possible behavioural dispositions and understanding any given phenotype requires understanding the individual learning process, but none of the individual learning processes is selected for. The capacity is – and whatever constraints and default trajectories it might involve. There is a continuity between simple one-or-the-other triggers that choose between two developmental paths (such as the reptile sex determination) and complicated general learning processes in evolved plasticity that involve information extraction from the environment and context-sensitive cognizing before internalizing a habit. The same logic applies all along the continuum. The degrees of freedom in the variation of the phenotype as a function of environmental influences is what *selection* works on, based on the consequences of allowing variation. Selection does not work directly on the phenotypic properties. It is, however, crucial to understand the developmental processes and how they result in the phenotypes in any given context in order to understand the *phenotypic evolution*. This is the core of the developmental expansion of the Extended Synthesis.[210]

How does the individualism and holism issue emerge, then? I have already fixed the assumption that replicators are individuals, but this does not mean that all *replication processes* are individualistic. I argued earlier that even if individuals are the interactors, some of their behaviour should still be understood as holistic, multi-individual

[210] To the extent that mainstream evolutionary developmental biology and the Developmental Systems Theory disagree on the unit of evolving trait, they frame the question differently. They are both right.

traits. I now propose a holistic interpretation of some replication (or reconstruction) processes. This is holism in a different dimension, and the traits that involve these processes do not need to be holistic traits in the proximate dimension, and the other way around. The idea is that some of the developmental resources in the developmental processes of some traits are shared by multiple individuals (either the same tokens or same types with shared history; I will expand on this shortly). The evolutionary relevance of this is that some developmental resources that are difference makers for phenotype and therefore fitness of individuals may be passed to the next generation collectively. The comparison is not between all factors but only those that exhibit variation within the population; sharing those elements in which there is variation in does. These elements may include collectively transformed or constructed environments (including technology and other artefacts that may influence the development) and horizontally transmitted traits (through different forms of social learning, but also the behavioural effects to the children's development that are similar to parental effects but caused by others). The core of the idea is that there is phenotypic variation between human individuals that are caused by developmental resources inherited from individuals other than parents (non-vertically) and these developmental resources (and their effects) are shared by other individuals too. If there is a systematic cofounder for sharing these resources, the replication process is holistic – not in the sense that the development of the individual is holistic, but in the sense that the development of the trait depends on the group-level reproduction or maintenance of the developmental resources.

If we approach development as a mechanistic process that made up of parts, the features of the developmental environments (including social and cultural environment) are shared parts of several individual developmental processes. The shared part does not need to be the same *token* of an environmental feature (any more than a shared gene needs to be the same token), but if two individuals share a same *type* of developmental resource and there is a *joint explanation* for why

this type of factor is shared (for example, the individuals share a culture), the developmental processes are *dependent on the same explanatory factor*. Mere externality of developmental resources does not make them holistic in any sense, of course. First, as stated above, the general features of the environment shared by the whole population do not count as inheritance with evolutionary significance. This requires difference-making between individuals in phenotypic variation. Second, if an external device for transmitting a trait to offspring, such as inherited materials or modified environment, behaviour copying, or direct teaching, takes place only on parental lines, it is merely a case of *external individualistic reproduction*.[211] Some such behaviourally transmitted external developmental resources may, however, exist (and spread) widely in a group and not exist in another group – these are cultural differences. The consequences for selection are direct: if the replication is individualistic, the selection is between individuals, but if there are shared resources, helping to pass that resource on through another individual may be selected. This shared resource may also be a "shared gene" – kin selection is a form of developmental holism. I will return to kin selection and its relationship with other cases of developmental holism in the last chapter. What is important now is that other developmental resources may serve a similar role of connecting two individuals reproductively – this includes modified environments, if there are differences between them among different populations, and culture.

We can distinguish between two senses of holistic process in the developmental dimension, just like in the proximate dimension. If

[211] Furthermore, a distinction needs to be made between selected non-genetic parental effects, whether they are epigenetic or some version of external transmission, and transmission of information that parents have recognized to be adaptive and pass on, whether this is conscious or non-conscious, and whether the route is symbolic, material or behaviour copying. The distinction between these categories (*selection*-based and *detection*-based) applies to different forms of cultural transmission as well (see Shea 2013); I will return to this later.

some individuals within the group (but not everyone) share developmental resources and are connected through this, this is *interactionist holism*. The individuals constitute a cultural subgroup through the resources; I call this "interactionist" since the paradigmatic case in the context of *social behavioural traits* is acquisition through interactions with others (through behavioural effects, copying, social learning or norm-enforcement, for example). If all the individuals of the group of individuals share a behaviour-shaping developmental resource necessarily, by virtue of being a member of the group, this is collectivist holism.[212] This takes place if there are shared cultural meanings and/or behavioural rules that all the members of the group share nearly universally because of identification, enforcement, or some other mechanism. Development without holistic elements is individualistic. This may include external routes of inheritance too; cultural transmission from parents to offspring is still vertical and works similarly to other vertical forms of transmission from the evolutionary point of view. Those characteristics that are not reproduced but are, for example, individually learned through individual experiences and not passed forward as such, are not directly objects of evolution, as discussed above. A special case of developmental individualism is evolutionary nativism, which considers only innate traits to be an object of evolutionary explanation.

The usefulness of the concept of innateness is contested, and the developmental individualism it assumes is incorrect. However, there

[212] Note that this notion of collectivist holism is weaker than the collective reproducers of Peter Godfrey-Smith (2009; 2014). Godfrey-Smith's collective reproducers are collectives in which some units of the collective reproduce while others only help to make this possible, examples being multicellular organisms (in which only sex cells reproduce) or superorganisms such as insect colonies (in which only queens and drones reproduce). Meerkats, among whom only alpha males and alpha females reproduce, are probably the closest mammalian example of this. In my sense of collective developmental holism, individuals reproduce but there are holistic elements that affect the development of every individual.

is a (slightly revisionist) way to understand the concept of innateness that makes sense and reflects the function of the concept on the fields it is used in. I will define and defend this in the next chapter. Nativism in evolutionary psychology, if understood this way, is not as much an empirical claim as a (developmentally individualist) methodological choice may be justified in some contexts but narrows the object of explanation in ways that the advocates of the methodology do not acknowledge.

The development of human psychology and behaviour lies on a continuum with the rest of the biological development, with an increasing amount of influence from outside and from other people. External influences on brain development start before birth. Psychological development takes place in stages with sensitivity periods, and different kinds of interaction are relevant to development in different stages.[213] The idea of generative entrenchment applies here too, especially the relevant early life interactions that are difficult to conceptualize as "learning", especially if the notion of learning is too intellectualist. A lot of early development is about practicing cognitive processes and emotional responses in interaction with the environment (including the social environment), which provides feedback that is often necessary for the development and sometimes instructive. This may include seemingly trivial things such as a mother's touch. As the child grows older, the specific features of the environment become more important in shaping cognitive skills, social emotions, and other capacities. Learning is a special case of plasticity; it is an extreme case of it rather than something qualitatively different. Learning involves cognitive processing of what is being learned, but to various degrees, and there is a continuum between brute causal influences and

[213] There is a slight discrepancy between phenotypic plasticity approaches and normal stage approaches, but both seem to be valid and there are ways to mitigate this discrepancy; see Love 2010.

symbolic learning.[214] The capacity to learn is a developing capacity itself and it is constrained and directed by the already achieved psychology. On the other hand, the developmental tendencies that seem to be species-typical and "pushing through" in development often rely on environmental factors. An important part of the interaction in psychological and behavioural development is the child's *activity:* children develop their "natural" skills and capacities by doing things, especially playing. Social environments (especially playing with other children) are important for cognitive, emotional, and behavioural development, and an important context for learning the social and cultural rules and meanings of the particular culture in which they are growing. (See Gauvain & Munroe 2012; Narvaez *et al* 2012; Berk 2013.)

7.2.3. Culture

Culture is a specific case of environmental influences. It is not the only thing that affects the development that we should consider a holistic factor from the evolutionary point of view, and I will not equate holism with culture. However, culture has a special role in thinking why human groups should be considered holistic units regarding behavioural traits, so I will discuss it in more detail.

Culture is also a very vague concept. "Culture" has multiple meanings even in the cultural sciences and it is not always clear when the term is supposed to be a classificatory term, an explanatory concept, or simply a category in general theoretical discussion about the nature of the research objects (see Kuper 1999; Fox & King 2002a; Ramsey 2013b; Koskinen 2014). Furthermore, the concept of culture that is used in biological contexts when studying whether an animal species has a culture or not, for example, has a different function and possibly a partly different reference (see Laland & Galef 2009; Ramsey 2013b).

[214] There is, for example, a difference between *imitating* behaviour (replicating it) and *emulating* it (abstracting its goals and reconstructing rather replicating it), which I will discuss shortly.

The main difference between the biological notion of culture that is of interest in questions such as "do animals have culture?" (Boesch & Tomasello 1998; Hunt & Gray 2003; Laland & Galef 2009; St Clair & Rutz 2013) or "how did (human) culture originate?" (Leaky & Lewin 1992; Boyd & Silk 2003; Boyd & Richerson 2005; Pagel 2012) and the one used in human sciences is that the biological concept of culture refers to the *medium* of cultural differences, and the various concepts of culture in human sciences refer to the contents within this medium: cultural systems, parts of them, or cultural entities. In the biological sense, humans have culture, in the singular. In the human sciences sense, humans have cultures, in the plural. I will not adopt any strict material definition of culture (such as specifying a form of learning) here. Since my perspective is evolutionary, I will follow Frans de Waal and Kristin Bonnie (2009; see also de Waal 2001) and characterize culture through its evolutionary function: horizontal social transmission of behavioural traits. The capacities involved, for example, are secondary to what is culture. The same capacities and processes may also be part of vertical transmission. But the nature of these capacities and processes matter.

Explanatory references to culture are mostly not proximate explanations, but shorthand for developmental explanations: the difference is due to the differences in the socially held beliefs, meanings and norms, as well as the material culture, that the people being compared have been exposed to. From the biological point of view, culture can be understood as something that comes with the "soft inheritance" (Mayr 1980; Jablonka & Lamb 2008), especially behaviour imitation and symbolic inheritance.[215] However, culturally transferred and

[215] This leaves two routes open for attempts to define (or characterize) culture from biological point of view: the reference may be this system itself (as it is in the context of human sciences) or the psychological capacity for culture. Culture as a system is usually characterized as information that is transferred from individual to individual (see Durham 1991; Heyes 1993 & 1994; Boesch & Tomasello 1998; Cronk 1999; Alvard 2003; Richerson & Boyd 2015; Ramsey 2013b; Paul 2018). Grant Ramsey, for example, defines it as "information

socially enforced *norms* and *rules* are especially important for social behaviour (see, for example, Sober & Wilson 1998; Boyd et al 2002; Fehr & Fischbacher 2004; Gintis 2006, 2007 & 2009; Roughley & Bayertz 2019). Enforcing norms through punishment or rewards is a proximate dimension activity that incentivises others to behave in some specific way. The rules that are enforced are, however, socially acquired. These rules are not necessarily *group-level* properties: there may be variation within the group in which behavioural strategy to choose and there may be competing normative expectations. But if the group enforces some norms across the group, if some norms become prevalent for some other reason, or if some cultural meanings for actions become part of the social reality of that group, these norms and meanings become the group's properties. This means that groups act as collectives in the proximate dimension regarding this trait. It may also enable collectivist group selection between groups based on cultural differences. For example, in Robert Boyd's and Peter Richerson's (1985 & 2005) cultural group selection model, these culturally created differences loop back to biological evolution by causing selection pressures for more efficient acquisition of these traits, which may result both in trait-specific selection and in selection for general forms of social learning, as well as other biological adaptations to the cultural environment (such as lactose tolerance in cultures that use milk).

---

transmitted between individuals or groups, where this information flows through and brings about the reproduction of, and a lasting change in, the behavioral trait" (Ramsey 2013b: 466). The psychological side of the issue is usually the related human capacity for social learning (for example, Galef 1992; Heyes 1994 & 2013; see also Call & Carpenter 2002; Call et al 2005; for criticism of equating culture with social learning, see McGrew 2009; Sterelny 2009; Ramsey 2013b). Yet others distinguish between cultures and a larger category of traditions based on the differences in the psychological mechanisms for transferring information (for example, Galef 1992; Whiten 2009).

Cultural differences in behaviour are instantiated in individual psychologies and how individuals interact.[216] They are learned from the social surroundings. There are various theories in psychology about how cultural features become internalized as part of psychology (see Zittoun & Gillespie 2015), but cultural differences in thinking styles, cognitive skills, emotional responses, and behaviour exist (Gauvain & Munroe 2012; Narvaez *et al* 2012; Berk 2013; Wang 2017). Participating in cultural and social practices and using cultural meanings and artefacts shapes psychological processes and the underlying brain processes (Kitayama & Park 2010), being a part of our biological developmental environment. Some aspects of culture, such as narratives and other representations, may function as entities that are copied from one mind to another, and the cognitive processes are just a selective environment for culture (Sperber 1996), which may still affect cognition and behaviour in radical ways[217], but at least some aspects of culture are developmental factors in psychological development. Furthermore, humans have the ability (and, to some extent, a tendency) to *imitate* others' behaviour whereas other animals *emulate* – humans do not only try to achieve the perceived goal of the behaviour, but copy parts of behaviour that would seem to be unnecessary or even pointless (Call, Carpenter & Tomasello 2005; Lyons, Young & Keil 2007; Whiten *et al* 2009; Heyes 2013; McGuigan *et al* 2017) and adults do this even more efficiently than children (McGuigan, Makinson & Whiten 2011). This serves as a starting point for cumulative culture and its evolution (Boyd & Richerson 2005; McGuigan *et al* 2017).

The difference between hard and soft inheritance is not a significant difference as such. There is no evolutionary difference between parents transferring behavioural traits through genes or behavioural

[216] The interactions may be governed by explicit (and sanctionable) rules of conduct, self-fulfilling expectations (empirical or normative; see Bicchieri 2006), or internalized patterns of interaction, acquired through learning by experience or through imitating others, for example.

[217] For example, some elements of religion can be explained in this way; see Boyer 2001 and Atran 2002.

and other external forms of transference. Culture does not *entail* horizontal inheritance. What is important, however, is that some of the forms of trait transfer are such that they can break parental linage. Even if they *can*, they do not necessarily *do* so. A trait that involves cultural transmission and becomes prevalent within a group may do so because every individual has got it from one of their parents. But it may be that "one of their parents" is exactly right – unlike in genetic inheritance where the acquisition of a trait depends on both parents, some forms of cultural transmission may be more efficient in spreading a trait. Take tool making for example. It may be enough that one of the parents knows how to make tools but exposure to this activity makes all offspring likely to practice tool making too. This could explain tool making becoming prevalent in the given population relatively rapidly without horizontal transmission. But in humans horizontal transmission is prevalent.[218] Cultural transmission does not make replication holistic, but its horizontality does.

Another important aspect of human culture is that it is not only a medium of inheritance and a developing environment; it is a selective environment. This aspect of culture has been the focus of *Niche Construction* theorists (see Laland *et al* 2000a & 2001; Laland & O'Brien

[218]Related to this, there is a distinction between selected non-genetic parental effects and the transmission of information that parents have recognized to be adaptive and pass on (Shea 2013). Some selected non-genetic parental effects are epigenetic and some use external transmission, including social learning, but they are still part of selected transfer mechanism. If the parents pass on information and behavioural tendencies that they have learned through their own learning, there is not only a different route of inheritance but a qualitatively different route for phenotypic evolution: behaviour is not selected naturally, but by individuals. The capacity for this is probably selected when it exists (see Tomasello 1999; Boyd & Richerson 2005; Baumeister 2005; Schaller et al 2010; Csibra & György 2011; Sterelny 2012), and there is probably overlap between the two processes (as, for example, in language acquisition). This type of transmission involves reliance on another individual's learning, which is the first step towards cultural transmission that is detached from individual learning.

2012): culture is part of the human-made environment ("engineered ecosystem") that humans adapt to, in the usual niche-constructing loop. This connects culture nicely to the general view of the extended synthesis: culture (or various parts of the complex that is referred to as "culture") plays various roles in the evolutionary understanding of humans. First, it is a part of the evolving phenotype (the psychological capacities that are its basis, the behavioural traits that we consider cultural, and material culture as the modified environments with artefacts and technology as the extended phenotype). Second, culture is a part of the developmental processes. It is usual to treat cultural processes as a further layer on top of individual developmental processes (for example, Boyd & Richerson 1985 & 2005; Sperber 1996; Jablonka & Lamb 2005; Odling-Smee & Laland 2011; Laland & O'Brien 2012), but this is only true of the cognitively higher parts of the culture: systems of belief, technologies, norms, *et cetera* that can be learned as an adult. The behaviour of others and the material culture (toys, for example) are, however, part of the developmental environment of children and can affect the psychological development of cognitive capacities, attitudes, behavioural tendencies, social skills, *et cetera*. Furthermore, the relevant socio-cultural environment goes beyond parents only in humans. (See Kitayama & Park 2010; Gauvain & Munroe 2012; Berk 2013; Narvaez 2014; Narvaez *et al* 2016; Wang 2017.)[219] Third,

[219] There are good reasons to think that some of the "channels" for cultural inheritance can (and should) be distinguished from biological channels; for example, Maria Kronfeldner (2021) argues this based on take-off autonomy (cultural creativity, change, and spread are independent of biological resources), near-decomposability (the interactions between cultural elements and between biological elements in development are stronger than cross-interactions), and differences in stability-granting temporal order (biological channels of inheritance convey more stability). This argument is mostly directed towards the Developmental Systems theorists (Griffiths & Gray 1994 & 2005, for example), who want to downplay the differences between different kinds of factors. They do not deny the existence of different kinds of factors, however, or some differences between them. Culture is not a homogenous

there is cultural evolution of its own in (probably most of) the various cultural entities.

Cultural norms of behaviour occupy all these positions too. Behaviour according to the norms, as well as potential enforcement practices, are part of the evolving phenotype, and there is evolution in norms. Imitating others' behaviour may be a part of children's development of behavioural dispositions that copies behavioural tendencies through ontogeny, and behaviour towards children may also result in other effects in children that enforce the copying of the norms: for example, gender-biased choice of toys and the encouraged and discouraged forms of play may enforce both gender-related norms and attitudes. Some cultural norms are replicated in early stages of development. The acquisition of some other norms takes place in more complex learning contexts, partly in adulthood, and are more strongly vertical in their spread. The cultural replication mechanisms of these norms may also vary. To mention a few: deliberate acceptance of a cultural norm, explicit or implicit acceptance of the norm as a way to do things and accomplish goals in the social context, copying a successful person or a person in an authoritative position (see Boyd & Richerson 1985 & 2005 for discussion of different strategies), submitting to enforcement, or non-conscious internalization of the norm. The enforcement of norms by others is a part of the cultural replication mechanism, and the emergence and maintenance of norms by social structures may be considered as such as well, if the social structures are partly constituted by culturally reproduced contingencies. And finally, the norms (both their existence and their content) are part of the selection environment for psychological characteristics.

The logic of cultural niche construction and how norms figure in it, and the integration of culture in the Extended Synthesis in general

---

category, however, and different elements (or even the same elements) may influence the development in different ways. These issues are not crucial for the main topic at hand – but that culture can affect development in different ways, and that there are different kinds of connections between culture and biological evolution, are important.

(call this *socio-cultural evolutionary developmental biology),* show that replication of behavioural traits within a group can be holistic, it can facilitate the evolution of interactive traits, and it can facilitate group selection by creating relevant group-level differences. It should, however, be just as clear that the existence of cultural transmission does not imply collectivism. This depends on the replication mechanisms involved, and it may depend on the mode of the cultural transmission, for example. Luigi Cavalli-Sforza (2000) distinguishes between four modes of transmission. First, inheritance through soft routes may be just as vertical as genetic inheritance, making them individualist factors in development. Second, horizontal modes of transmission include *one-to-one* mode, *magistral* (or one-to-many) mode (such as copying from an authority figure), and *many-to-one* mode (such as the social group enforcing a norm). Horizontal transmission may result in similarity between some interacting individuals within the group (interactionist holism) more likely than uniformity across the group (collectivist holism).

Furthermore, not all holistic developmental processes are as easily evolvable as the simple cases of individual beliefs or skills. If the development relies on resources provided by others and these resources are reliably available during evolution, the development might become fixed to some degree. Take Darcia Narvaez's idea of the *evolved nest* (or *evolved developmental niche)* for example (see Narvaez *et al* 2012, 2014 & 2014). According to her, there is a combination of behavioural effects (provided mostly by the parents, but other individuals of the society are also an important part of the "nest") that development relies on during various stages of development. A lack of these developmental resources skews the development of emotions, cognitive capacities, and attitudes. These resources were present in the EEA, and the *selected outcomes* only appear if they are present now. Language acquisition may be a similar example: languages are cultural, but they involve the language-specific cognition that develops only in the interaction between the psychological developmental tendencies and the linguistic environment. Many sociality-related

traits may be like this in their development: socio-cultural factors are enabling but not instructive.

What consequences does this all have for evolutionary explanations? Culture facilitates both interactionist traits and group selection, but neither is directly entailed. There is a temporary dimension in evolutionary processes involving culture. Culturally specific social behaviour can be understood as a product of recent evolution in evolutionary anthropological research. There is, however, a difference between a "cultural trait" being a trait of a group or spreading within the group culturally. These have different implications for whether we should expect to find holistic or collectivist traits under the evolutionary analysis, and for the levels of selection issues. But as noted above, the mere presence of cultural processes does not imply any of this directly. Furthermore, the existence of cultural factors in the development of a trait does not mean that the trait could be an object of cultural evolution alone. As for evolutionary psychology, culturally specific behaviour is not something that it is usually interested in at all. However, the capacities for culture have evolved in a social context and their function may be a part of an interactive trait (as I argued in the previous chapter), and many social capacities that evolutionary psychology is interested in might require a social developmental environment. Neither is an argument against evolutionary psychology, but it affects how to do it properly. On the other hand, just because a capacity develops in interaction with its environment does not necessarily make nativist methodology inadequate. I will go into this in the next chapter.

8. Innateness and Nativism

The previous chapter discussed some of the reasons to take the details of development seriously in thinking about evolution, as well as some reasons to consider how shared external developmental resources may connect individuals through the developmental processes in ways that are relevant for evolution. I will discuss this in more detail in the next, final chapter of the dissertation. However, it is not clear what developmental holism means if it is not properly contrasted to what the alternative, developmental individualism looks like. My aim is not only to point out the well-known fact that development and its environment matters in the Extended Synthesis framework, but to make the distinction between individualist and holistic developmental processes in a way that matters for evolutionary explanation. What does developmental individualism, which implies developmental isolationism, even look like within this framework? Furthermore, although (some of) evolutionary anthropology studies contemporary societies, and their cultures as part of the adaptive functionality of those societies, evolutionary psychology (in any of its forms) is interested in features of mind that are not culture-specific, or environment-specific in any other way. Does Extended Synthesis extended with culture (which I called "socio-cultural evolutionary developmental biology") simply entail that this is impossible? My answer to both is: there are innate traits that are the object of developmentally individualist evolutionary explanations.

The concept of innateness used to be a central concept in controversies over biological (especially evolutionary) approaches to psychological traits and human behaviour. The distinction between the biological and cultural was often equated with the distinction between innate and learned, and many scholars considered only biological traits to have evolutionary explanations. This picture has become more complicated in three ways: (1) The clear-cut distinction between innate and acquired has been put in question in biological contexts. Furthermore, some have even considered the very concept of innateness to be

confused and unsuitable for scientific purposes in the first place (Griffiths 2002 & 2009; Griffiths & Machery 2008; Mameli & Bateson 2006; Bateson & Mameli 2007). (2) Doubt has been cast on the usefulness of biology-inspired notions of innateness in psychological contexts. The proposed notions (for example, innateness as canalization (Ariew 1999 & 2006)) do not seem to reflect the function of the concept of innateness, for example, in the language acquisition debate (see Griffiths & Machery 2008, Kiikeri & Kokkonen 2007; O'Neill 2015). (3) The priority of both genes and biological innateness in understanding evolutionary process has been challenged, especially by advocates of the Developmental Systems Theory, as discussed in the previous chapter, and gene-culture co-evolution theories emphasize culture in their evolutionary explanations of human behaviour (Boyd & Richerson 1985 & 2005). In the light of these considerations, innateness seems to have a lot less to do with evolutionary approaches to psychology or behaviour. On the other hand, the idea of nativism occupies a central place in cognitive sciences (Samules 2002, 2004, 2007 & 2009; Kiikeri & Kokkonen 2007; O'Neill 2015) and it seems like a natural starting point for evolutionary approaches to psychology (in contrast to behaviour).

Much of evolutionary psychology is nativist. This is a form of methodological individualism in the developmental dimension: development, when it comes to explanatorily relevance, is independent from others. Developmental holism, as an alternative, assumes connectedness to others through developmental resources. Individualistically understood developmental processes include individual learning, but the contents of individual learning are not an object of biological evolution. The capacity to learn (and what is learned and how it is learned) is. The fixed features of development, including learning capacities (if not the results) are what we tend to call "innate". Individualism and nativism are not co-existent: individualism applies to both innate and individually learned traits and some innate traits have an evolutionarily meaningful reproductive dependency on others (for example, through kin selection). We can treat the latter as a special case with additional criteria (I will return to this later), and the

evolutionary considerations here will be about the innate traits. Therefore, nativism and individualism can be equated most of the time. How nativist evolutionary psychologists approach their research makes them pure individualists along this dimension.

In everyday thinking, the distinction between innate and acquired is usually thought to be both valid and mostly clear-cut. Furthermore, it is often equated with the distinction between biological and non-biological. This used to be the scientific view on the issue, for example, in ethology[220], and continues to be shared by some. The complications for the distinction have however, been acknowledged by biologists, also by evolutionary ethologists since at least the 1960s[221], and the uselessness of the concept of innateness is certainly already a truism among developmental biologists and philosophers of biology. Some contemporary philosophers have even argued that the whole concept of innateness is confused and not suitable for scientific purposes at all.

However, the distinction continues to be an important theoretical presupposition for some evolutionary psychologists and a heuristic for many other psychologists (see Samuels 2004; O'Neill 2015), a situation which, minimally, calls for scrutiny regarding what they – or those who disagree with them – mean by the concept. Furthermore, it is certainly a part of the public understanding of evolutionary approaches to human beings, leading to false conclusions and misunderstandings. For some, this is a further argument to abandon the concept of innateness altogether, but it could just as well be an argument to clarify a more scientific notion of innateness, as I will be doing. There are some cognitive scientists (probably the majority of them) and philosophers of psychology (for example Khalidi 2002 & 2007; Mallon & Weinberg 2006; Samuels 2002 & 2007; Weinberg & Mallon 2008) who

[220] There used to be a somewhat clear distinction between innate *instincts* and learned behavioral patterns in the study of animal behavior. For more about the historical idea of innate instincts, see Griffiths 2004 and Brigandt 2005.

[221]See especially Hinde 1968 and Tinbergen 1963. Even Konrad Lorenz, the most prominent defender of the concept of instinct, blurred the clarity of the distinction (see Lorenz 1965).

think that this can be done in psychology: there is a scientific concept of innateness to be explicated and some sort of nativism is defendable, even required, for psychological theory, and they use it successfully as an assumption in research. This raises the question of whether the notions of innateness criticized in biology and defended in psychology are even the same (see Kiikeri & Kokkonen 2007; Samuels 2007; O'Neill 2015). If not, a corollary question emerges: which notion of innateness is used in the nativist evolutionary psychology?

I will not try to evaluate the validity of any form of nativism as such, evolutionary-psychological or otherwise, in this dissertation. I will, instead, focus on the issue of the *meaningfulness* of the concept of innateness, its definition, and the role it plays (or should play) in explaining individual development and in evolutionary explanations, and in how these are connected. Nativism about any particular trait as such is an empirical issue and therefore outside my direct discussion, although I will use some examples (all of them controversial, to be sure) to explicate *what it would mean* if a trait were innate. There are two sides to this issue. First, whether there is a scientifically meaningful notion of "innateness" at all. I will argue that there is. Second, whether it makes an explanatory difference whether a trait can be called innate or not. I will argue that it does, *especially* from the evolutionary perspective. I will, perhaps surprisingly, defend the nativist methodology in evolutionary psychology, framing it as a choice of explanatory questions. It is, however, important to distinguish between the dimensions. Even if evolutionary psychology were only interested in the innate (psychological) parts of the social trait, the (behavioural) trait itself could be an interactionist trait, which determines its evolutionary function, and this could be a product of group selection. None of this is determined by nativism, and these explanatory dimensions should inform even nativist evolutionary psychology.

In the first section, I will discuss why some philosophers (and others) think that the concept of innateness is fundamentally faulty. I will discuss the folk notion of innateness and the difficulties it faces when applied to real biology. I will, however, argue that it makes

sense to try to find a meaningful notion of innateness and criteria for the definition to be successful. In the second section, I will build a working definition for the concept, a *contrastive invariance account* of innateness. I will also distinguish between biological and psychological concepts of innateness. The concept of innateness has (or should have) different references in biological and psychological contexts, although the explanatory role of the concept might be similar in structure and the developmental process referred to can be the same. I will also argue that the criticism against the usefulness of the biological concept and the defence of the usefulness of the psychological concept can coexist peacefully. Finally, I will return to individualism and holism issue and draw conclusions from my discussion for nativism debates in evolutionary psychology and elsewhere.

8.1. What Is Wrong with Innateness?

The distinction between innate and acquired cannot be as clear-cut as it is thought to be in everyday thinking, for many reasons. Some philosophers and scientists have, however, gone further to claim that the whole notion of innateness is just that – a part of everyday thinking. It is part of *folk biology*, a proto-scientific way of understanding biological reality (see Atran et al 2001; Atran & Medin 2008), and it has no use in science. For example, Paul Griffiths and his collaborators Edouard Machery and Stefan Linquist (Griffiths 2002 & 2009; Griffiths & Machery 2008; Griffiths, Machery & Linquist 2009) have taken this line of argument. According to Griffiths *et al*, when the concept of innateness is moved from the folk-biological framework to a scientific context, the connotations follow, with confusing results. Others, too, have pointed out that the use of the term "innateness" seems to be too vague for scientific purposes. Matteo Mameli and Patrick Bateson (2006; see also Bateson & Mameli 2007 and Bateson 1991), for example, have systematically discussed as many as 26 different properties that are or could be linked to innateness, none of which they find

satisfactory as a definition, and, furthermore, are not necessarily connected to each other. The argument following from these considerations is an *eliminativist* argument: the concept of innateness is confused, it does not refer to anything substantive in biological reality, and therefore has no real use in biological sciences. In the following I will discuss the eliminativist challenge in greater detail and try to specify exactly what is wrong with innateness and see what can be saved. I will start with the idea of innateness as a folk-biological notion and what can and cannot be inferred from the fact that it is one. I move on to explicate some of the problems with the concept in biology and then to the various strategies proposed to save the concept.

8.1.1. Innateness as a Folk-theoretical Concept

As already discussed in the context of folk psychology, human beings have sets of "common sense" ways to deal with reality around them – ways to conceptualize what they perceive, ways to make inferences, predictions, and so on and so forth. These commonsense ways, *folk theories*, include substantial but usually non-explicated presuppositions about the nature of their subject matters. At the heart of folk biology is its *essentialism*: species, races, and other taxonomic levels reflect how individuals are bound together by their shared essences; the shared essence explains why all the members of the species (and other groups, below and above the species level) tend to possess certain traits; and these essences can be used to make generalizations and predictions across the species and other folk-taxonomic levels. (See Medin & Atran 1999 & 2004; Medin & Ortony 1989). According to Griffiths (2002), innateness is best understood as a quasi-theoretical concept central to folk biology. The essential traits (perceivable properties caused by the unperceivable essence) are primarily thought to be innate, and this essence is the source of innateness. The traits have also a *function* derived from the essence – they have a purpose from the point of view of the organism's overall design. More specifically, innateness

is connected to three different ideas: (1) *Fixity*[222] – the outcome of the development of the trait is fixed in individual development and it is hard to change through external intervention; (2) *Typicality* – the trait is universal or at least typical among the members of the species; and (3) *Teleology* – the trait has a purpose or a function (which translates to the evolutionary framework as being an adaptation). The problem is, for a start, that the essentialist way of thinking (or typological thinking in a broader sense) about species (or biological entities in general) is fundamentally faulty[223], and to be more precise in this particular case, the connotations noted above are not necessarily connected in the real world.[224] This leads to imprecise use of the concept of innateness and erroneous inferences when it is used in scientific contexts.

Folk biology and its development in children has been studied for some time, and the essentialist nature of folk biology (including thinking about human ethnic groups) is well known (see Atran 1990; Atran *et al* 2001; Atran & Medin 2008; Berlin 1992; Gelman 2003; Gil-White 2001; Hirschfeld 1995; Keil 1989; Medin & Atran 1999 & 2004; Medin & Ortony 1989).[225] Some external, perceivable properties of an organism

---

[222] I use the terminology from Griffiths, Machery & Linquist 2009.

[223] For differences between essentialism and typological thinking on the one hand and biological population thinking on the other, see Futuyma 1998; Mayr 1959; Sober 1980.

[224] Mameli and Bateson (2001) go even further and distinguish eight so-called "i-properties" that do not go hand in hand; I will return to this later.

[225] It should be noted that the word "essentialism" is used here in its psychological meaning, referring to a substantial way of thinking about certain entities. Much of the philosophical discussion about essentialism has been on the distinction between essential and accidental properties and whether some objects have essential properties that are necessary for them to be those objects. This may lead to quite minimalistic views. In philosophical terms, it might be more correct to say that *psychological essentialism* is the tendency to presuppose various classes of objects being *natural kinds*. Insofar as the tendency to attribute underlying essences to, say, animal species, does not depend on observable properties but explains them when they exist (as the experiments with children seem to suggest; see Keil 1989 and Gelman 2003), it might not even be necessary for these natural kinds to be *essentialistic* natural kinds (Ellis 2001), just a propensity

are thought to be caused by its inner nature, while others may be caused by the environment. Even essentialist external properties may be subject to influences from the environment (for example, under a direct causal intervention), but they are not thought to develop in *interaction* as in (a real biological) gene–environment interaction; they can only be distorted. Furthermore, the essence stays unchanged by the environment and is permanent. For example, even relatively small children think that (1) the species-typical appearance of an animal comes from within, (2) it *can* be changed by an external intervention to resemble the appearance of another species, but (3) changing does not affect its other species-typical properties (Keil 1989). Nor do environmental changes (such as being raised by animals from another species) change an animal's species-typical behavioural dispositions (Gelman 2003). The essence is related to external properties in a star-shaped causal structure (that is, each property is caused by the essence independently; see Kokkonen & Pöyhönen 2012), and environmental influences can disturb some of these causal connections at most.

The organism does not have only one essence; there is a (non-overlapping) taxonomic hierarchy of essences. The core level is what Douglas Medin and Scott Atran (1999) call *generic species* (which may coincide with a scientific species, but often coincides, for example, with a genus), but an individual organism has, for example, the essence of its race (or another sub-species group) and broader kinds, such as animal or plant, and a sub-group of animals or plants, and each essence has a range of properties associated with it (see Atran 1990; Berlin 1992; Medin & Atran 1999). At least some non-group-distinctive subgroup differences such as personality types may also be included

for properties to cluster when the underlying essence (in a vague sense) is present. On the other hand, this vague idea of essence includes the idea of *true* essences of things (*contra* natural kind ideas after the fashion of Dupré 1993). However, the precise philosophical characterization of folk essentialism is of no real consequence here. The main point is that the folk essentialism sees some traits as being (purposeful) consequences of an underlying, kind-shared, fixed essence that has explanatory and predictive force.

as part of psychological essentialism in addition to group-distinctive differences like those between ethnic groups (see Prentice & Miller 2007; Spears *et al* 1997; Yzerbyt *et al* 2001). Categories based on these essences are used in generalizing and explaining and predicting behaviour. Some levels are more important than others (the species level being clearly the most important), but they all exist, enabling, for example, inferences across species based on "essential" commonalities, instead of commonalities in form and function. For example, ostriches are thought to be more similar to owls than bats are, based on the knowledge that ostriches and owls are both birds, although bats and owls share more in their ways of life. The problem with this all is, of course, that the biological reality does not work like this.[226]

The idea that vernacular use of the concept of innateness has something to do with folk biology and psychological essentialism seems plausible enough and I will not dispute it. In the Griffithsian view, innateness is connected especially to *species level* essences (Typicality as species-typicality), as well as with Fixity and Teleology. Griffiths, Machery & Linquist 2009 empirically tested the idea that Typicality, Fixity and Teleology are three dimensions that people associate with innateness. They tested people's intuitions with various combinations of these dimensions in the development of singing in birds and discovered that people use them all in deciding whether a trait is innate or not, although Fixity seems to be the strongest and Teleology

[226] See Wilson 1999 for an overview of classic problems as well as some attempts to save biological species as a natural kind concept, even if not an essentialist one. Non-essentialism of species is foundational for evolutionary theory (see Mayr 1959; Sober 1980; Futuyma 1998; Gould 2002), and although homologies are an important phenomenon related to higher taxonomies (see Griffiths 2006; Ereshefsky 2007 & 2012), locating a species in a higher taxon has limited explanatory consequences. Furthermore, many everyday classifications are erroneous. For example, fish, wasp, reptile, crow, and tree are all polyphyletic classifications. There might also be pragmatic reasons to use different classification systems for different purposes in biology (although this is not the practice at the moment), leading to ontological pluralism (see Dupré 1993).

the weakest predictor for the judgment. Their later studies fail to associate Teleology with innateness at all (Linquist *et al* 2011), so this connection will be best left open, but I will include it in the discussion, just in case.[227] Since people connect all these ideas to innateness, they will also connect these ideas together and make inferences from one to another. The conclusion is that an adequate analysis of the folk concept of innateness needs to include all these dimensions but also some idea of their relative weights and interaction. They also make a more important point that both philosophers' and scientists' use of the concept of innateness probably also has a lot to do with folk biology – folk-biological categories guide thinking even in some scientific contexts, just like folk-psychology does, as I argued earlier. The problem with this is that folk biology is *wrong* about biological reality in this. Typicality, Fixity and Teleology are empirically separate (and there are also evolutionary reasons for them to be separate; I will discuss this later in this chapter). This entails the main point, the eliminativist stance on innateness: the concept is always the folk concept, with no reference in reality, and its use in biology, psychology and elsewhere should be stopped.

I agree with most of the above. Folk biology is likely to be an important source of confusion and vagueness in the use of the term "innateness" in biology and sciences in general (perhaps resulting in the many unnecessary uses of the term as argued by Mameli & Bateson 2006 and discussed in the next section), and philosophical intuitions about innateness are probably mostly the products of applying folk biology.[228] But I have a few objections to Griffiths and his collaborators.

---

[227] In fact, they find a connection between Teleology and *being genetic*, even without the connection between Teleology and innateness. This is somewhat puzzling, if being innate is being a part of essence, and genes are interpreted as essence-bearers in everyday thinking. However, this may say more about the superficiality, plurality of ideas, and lack of coherence of folk theories in general that was discussed in the context of folk psychology in an earlier chapter.

[228] The most revealing (although anecdotal) example of pure folk biology in action in philosophical analysis must be Devitt 2008.

First, it is one thing to point out that the folk theories of individual development are unscientific and even substantially contradictory to what is scientifically known. It is a completely different thing to claim that a central concept of these theories has no reference. The claim is, of course, not as simple as that. According to Griffiths *et al*, the concept of innateness is a conceptual connector of unrelated ideas, a cognitive structure that guides the use of the term "innateness" and connects it closely with the ideas of Fixity, Typicality and Teleology. But this can be disputed. Consider the following alternative interpretation. The folk concept of innateness refers to (developmental) Fixity alone (this is the strongest correlation), but folk biology also includes ideas about what traits are *likely* to be traits of this kind. In other words, there are criteria for judging a trait as innate or not based on the *substantial presuppositions* of folk biology. For example, folk biology presupposes essences (or "inner natures") that are universal within the species (and to some degree within some other "levels" of folk taxonomy), and obviously, essential properties must be inborn, not acquired, since the essence is shared and internal, not caused by external similarities in environment. Species-typical traits *must* be developmentally fixed. The connection between Fixity and Typicality is not conceptual, but a substantial claim about common cause for them that enables inferences. Folk biology also sees nature in general as purposeful (the properties of an organism are *for* something), and so Typicality gets bound with Teleology: whatever the species-typical traits are, they must have a function for that species. Typicality entails both Teleology and Fixity.

Since the examples used by Griffiths, Machery and Linquist (2009) in their survey are already about how different bird *species* develop singing, people use their intuitions about *species essences* in the test. This might be adequate for studying folk biology in general but inadequate for studying the concept of innateness in particular, or at least to test the hypothesis that innateness is bound to the idea of Typicality (on the species level). Even leaving this aside, if most traits that are fixed in development are thought to be species-typical – given that

the species-level is the most essential level – Fixity, Typicality and Teleology should substantially entail each other to some degree. But this is about the substance of folk biology, not about the meaning of the concept of innateness. I would propose that the core idea of even the lay concept of innateness is simply the idea of developmental fixation, and the notion is perfectly compatible with properties that are neither species-typical nor functional. This is, of course, an empirical claim that I will not defend here as such.

The meaning of the vernacular concept is also beside the point if our interest is in whether there can be a scientifically useful concept of innateness. But there are two relevant issues here. First, eliminativists propose that folk biology biases how scientists (and philosophers) think about innateness within science. This is a potential biasing factor to be taken seriously – especially since I have voiced similar worries about folk psychology myself in an earlier chapter. Secondly, especially if the first point is correct, the use of the term "innateness" in the first place must be justified: the lay concept must be similar enough to the scientific concept. I take the results of Griffiths *et al* to be evidence for this point: Fixity has the strongest correlation with being innate, and if my interpretation of the rest of the results is defensible, they have not demonstrated the necessity of the inclusion of the other dimensions in the meaning of the concept directly.

Furthermore, why should one focus on *species*-typicality in the first place? True, it is the strongest level of essentialism according to the studies, but not the only one. People use both higher and lower taxonomic levels to make similar inferences. Higher taxonomic levels could be interpreted as similarity-groupings of species-level essences, but the levels below species cannot exhibit species-typicality. For example, differences between ethnic groups are sometimes (erroneously) considered to be (generic) innate differences. Other, non-taxonomic examples of sub-species group-level innate (perhaps essential)

differences could be personality types and different learning styles.[229] This is, however, easily correctable. Griffiths *et al* could expand Typicality to encompass other levels as alternatives – a trait could be species-typical, bird-typical etc. But what does not fit their scheme is innateness on the individual level. It seems that people speak of individual-level innate differences in capacities, innate diseases, and so on. These differences between individuals are assumed (even if often falsely) not to be acquired. Person-typicality as a notion seems like a stretch, although a person's innate capacities, behavioural tendencies, etc. are undoubtedly typical to them. Species-typicality does not, of course, require *monomorphism* (for example, sex-differences are species-level properties) and the range of variation in differences between two individuals (for example, in hair-colour) is species-typical as such. But the difference between two individuals is not a population-level trait and being, for example, exceptionally talented in music is an innate characteristic of an individual in vernacular use of the concept.

Most scientific debates that have something to do with innateness are about species-typicality (such as nativism about certain features of natural language), but the idea of innate properties of an individual – personal but not acquired characteristics – would seem to accord with the vernacular use of the concept.[230] But even some scientific debates are about whether specific differences between individuals such as sexual orientation or gender identification are innate or acquired. Even if the range of possible orientations is considered a species-level

---

[229] It may be the case that neither personality types nor learning styles really exist, but they are *attributed* to people as groups and they are thought to be internal, inborn, not learned, and virtually immutable.

[230] Languages differ in how they express individual inborn differences. For example, in Finnish, the equivalent of the English phrase "natural born" (e.g. "natural born violinist") is "synnynnäinen", which means innate. A "natural born violinist" is not, of course, born with the related skills but a capacity to acquire them efficiently and a potential for reaching a high level. The logic behind the idea is still the same: there is a set of innate capacities distinguishing the person from others that makes this possible.

property, the debates on innateness of sexual orientation and gender identification are about individual-level differences and individual-level factors. If the species level is abandoned, the idea of Teleology also becomes more problematic. Biological species are thought to be functional in their environments, but individual innate properties such as innate diseases can be dysfunctional just as well.

In other words, I would dispute the idea of Typicality and especially Teleology being *conceptually* connected with innateness in folk biology. I would say, rather, that the core idea of innateness is Fixity. This is also the main focus of all the accounts of innateness that I will discuss shortly. Folk biology, however, has an *essentialist theory about fixity*. The connection of the three ideas lies in the *substance* of folk-biological ideas. We should distinguish between the folk concept of innateness and the folk explanations of innateness.[231] Of course, all this is relevant only to the justification of the use of the vernacular word "innateness", whether something like Fixity can be defined and found to be a useful concept and, more importantly, whether this concept is of valid scientific use. This does not, however, mean that the exact folk-biological concept of innateness can be *explicated* in terms of modern biology (or psychology). What I will do in the next subchapter is a minor *revision* of the concept in a way that fits the biological reality. The danger of revisionist projects is that the meaning of the concept can change too much in the process for it to really be the same concept anymore – it becomes a stipulation with a familiar name. A successful revision should give a more precise meaning for the concept, one that is more projectable to the subject matter, but not change the core content and the reference altogether. This has been done in

[231] This is similar to the compatibilist take on the notion of free will. The concept of free will has certain functions in moral practices, and there is a folk-metaphysical theory that explains why people have free will. Compatibilists take the concept of free will (with its function) but replace the folk-theory with some determinism-compatible account of what it means to have free will. (See Dennett 1984; Kane 2002; Mele 2009.)

science successfully, when our knowledge of, for example, physical reality has increased.

Let us take a closer look at the example that Griffiths *et al* (2009) mention as an analogy: heat and a snow-shovel with wooden and metal parts. According to folk physics, *heat* is a one-dimensional quantity: it is something that an object has more or less of, it is capable of warming other things up, and it is perceived as the hotness of the object. So, for example, if the metal part of the shovel feels colder, it should have less heat. We now know that what was considered heat (and still is in folk physics) is a much more complicated phenomenon.[232] First of all, it is not an independent quality, but reducible to the movement of the particles of an object. The more kinetic energy is present in this movement, the more *thermal energy* that object has. The mean kinetic energy of these particles defines its *temperature*. Two objects may have the same temperature but a different amount of thermal energy, depending on the structure of the substance that the object is made of. If objects are in connection, there is a transfer of energy from the hotter to the colder as long as there is a difference in the temperature, but the amount of thermal energy does not spread evenly. *Heat*, in its proper physical sense, is the energy that transfers from one body to another through contact or thermal radiation. If the object has a higher temperature than its surroundings, the more thermal energy it has, the hotter it is (in the same surrounding temperature) – this is a matter of thermal energy, not temperature. Finally, there is the phenomenological aspect of heat: how it feels. This, of course, depends highly on the actual physical heat. But in the example of a snow-shovel, which is not hot but cold (we are out in the winter frosts now), all these aspects are separate. The metal part of the shovel feels colder (it sucks heat from your hand more aggressively) – from a phenomenological point of view, the wooden part is "hotter". The metal part, however, has more thermal energy, so it is "hotter" in this sense. The temperature is the same. In both cases, the heat proper is negative (the

[232] For more about the scientific use of word "heat", see Brookes *et al* 2005.

shovel has lower temperature than hand) – so in a sense the wooden part actually *is* "hotter", but you cannot make any inferences based on that which would depend on the properties of temperature or thermal energy, which is the core metaphysical property that the folk concept of heat might be considered to be attempting to refer to in most cases.

So, it seems that the folk-physical concept of heat turns out to be four different things: *thermal energy* (the prior physical property, reducible to the more fundamental property of kinetic energy), *temperature* (a property derived from thermal energy and its behavioural tendencies related to the material properties of the object), *heat* (in its proper physical sense, which is also a derived property), and phenomenological heat. This does not mean (and Griffiths *et al* do not suggest that it does) that the folk-physical concept of heat lacks even a vague reference, or that the scientific use of the concept should be stopped, or that the folk notion and the scientific notion are trying to refer to totally different things. It is more correct to say that the scientific understanding of heat has revealed the phenomenon to be more complex than presupposed by folk physics, that folk physics gets things substantially wrong, and that the folk notion of heat succeeds in referring to only one aspect of the phenomenon it tries to refer to (heat proper). There is no reason to eliminate the concept of heat, just to update the folk understanding of heat. This is precisely what I will propose with innateness. There is a way to define "innateness" scientifically and it is close enough to the folk-biological concept for the same word to be used. Furthermore, the debates on psychological nativism are *mostly* about this property, although the connotations driven by folk biology may lead intuitions and inferences astray. But folk biology gets the underlying biology wrong. Innateness is the fixity of the outcome of a developmental process (the *core* meaning of the folk concept, just as heat proper is the core meaning of the vernacular concept of heat), under certain conditions. The folk-biological understanding of the underlying mechanisms (the biological essence) is wrong and leads to incorrect conclusions (just as folk physics gets the physics of heat

wrong).[233] In the following, I will discuss possible obstacles to this approach.

8.1.2. Problems for a Scientific Concept of Innateness

There are two problems that the scientific concept of innateness needs to handle. One of them is in continuation with what is discussed above: the concept of innateness seems to have several meaning components in the sciences as well. The properties or processes that are referred to as "innate" are not substantially connected in biological reality. Another is that the genes (and other internal developmental factors) and the environment of an organism interact in too complex and orchestrated a fashion for a meaningful distinction to be made about what traits are caused by internal factors and what are caused by the environment. As long as innateness is thought to be something that comes "from within", it is simply a meaningless concept. In the following, I will detail these issues. In the next section of this subchapter, I will outline some of the strategies that try to avoid these problems and in the next subchapter I will follow the line that I am taking myself.

---

[233] There is a crucial difference between the cases: the concept of innateness has social consequences; the concept of heat – not so much. This is also a point frequently made by the eliminativists, but I would question the effectiveness of the strategy of dropping difficult concepts. If the concept of innateness (with all its confusions) is going to be a part of folk-biological understanding anyway, whether the actual word is used or not (for example, in a popular science books), it would be more useful to define the concept properly and be explicit about the connections between innateness and other biological properties than to just drop the use of the term or, even worse, say that there is no such thing. A folk-biological reading of such a statement would be that this is a substantial (and a radical) statement about the *extent* of innate traits (namely that there are none), not about the dichotomy of innate and acquired. This is, however, another matter. The main issue here is whether a useful scientific concept of innateness can be defined in the first place.

Let us start from the simplified, lay-based distinction between innate and acquired. As discussed in the previous section, I will equate innateness only with developmental fixity for the present purposes. At the heart of this distinction, between innate (or inborn) and acquired, is the idea of some things coming from within, being already present at birth, and some things appearing because of the environment. The presence of a trait at birth is does not have to be concrete in this distinction: it is something that is waiting for appearing in the normal course of development without any guiding external causal contribution. For example, adult characteristics that appear during puberty are still innate. Neither is it necessarily fixed in the sense of being not changeable. For example, things like innate speech defects or postural anomalies are considered innate even if they can be, and indeed are, actively changed. The point is that they are not caused by the environmental causes, but that the "normal course of events", without direct intervention, is somehow determined by what is "within". There is also an idea of *partial* innateness even in the vernacular understanding: being innate can come in degrees. A trait being partly innate means that the development of the trait cannot be explained by external factors alone. A reference to something internal is also needed, and this "something" is conceptualized as being innate.

It therefore seems that the crucial difference is between what comes from within and what comes from outside the organism in the development. An obvious idea to look for innateness is to look, then, for what this "something within" is in development, and the obvious hypothesis would be equating innateness with "being genetic." However, this fails just as obviously, for the reasons discussed in the previous chapter. Genes are not the only internal factors in the developmental processes: a gene plays a causal role in development only if the environment (internal to the organism) is such that the gene will be activated as a part of the causal process and the exact way in which it contributes to the process is determined by other factors (including other genes indirectly), and so on. There is no direct connection between genetic "information" or "encoding" that could be simply

equated with innateness[234] and the phenotypic traits that are the end-product of a developmental process. (Godfrey-Smith 2000; Jablonka & Lamb 2005; Wolpert *et al* 2010.) Especially if we take the Developmental Systems Theory approach, genes lose their specific status, and the distinction between internal and external loses its meaning as well.[235]

Being merely "genetic" would not be enough anyway, if being genetic means only that genes play a role in development. Specific environmental factors could still be necessary and the outcome of the trait's development could be highly sensitive to different environments to the extent that talking about innateness would be absurd. Additionally, no form of learning is without some contribution from genes, even if only in the development of the capability to be sensitive to the relevant factors in the environment. At first sight, a tempting solution might be to introduce *degrees* of innateness that reflect the graduality of being genetic. There is, after all, a measure for this: heritability. (See for example Horvath 2000 for this suggestion.) But heritability is a population-level, not an individual-level, notion. It is a local measure that expresses the extent to which the (actual) variation in the genome results in the variation in the phenotype in the given range of environments. It also depends on the actual developmental mechanisms, but the value of heritability can change if the range of environment changes or if the genetic variation changes, regardless of the developmental mechanisms being the same on the individual level. It just is not a quantitative measure of the causal contribution of the genes on a population level under specific conditions. And even if we had a way to know, for example, that 60 % of my height is due to

---

[234] It is not difficult to get the impression that innateness and "being genetic" is *also* often equated in the lay understanding of biology by many psychologists and philosophers (see e.g. Buss 2003; Chomsky 2000; Fodor 2001; Pinker 1998a & 2002; Plotkin 1997; and Tooby & Cosmides 1992 for characterizations in genetic terms). Sometimes an exact equation is made explicitly: e.g. Fodor *et al* (1974: 450) states straightforwardly that "Language is innate *iff* it is encoded in the gene code."

[235] This is, of course, a part of the background of the eliminativism about innateness.

my genes and 40 % is due to my diet in childhood, it would not mean that the genes grew 110 cm of me, and the rest was built by eating. (Lewontin 1974; Sober 1988b, 1998.)

The interaction between genes (and other internal developmental resources) and environmental factors is non-additive in an even more radical way: it could be necessary for the development of a trait to have both a certain genetic makeup and a certain developmental environment for the trait to develop in the first place. This is the case especially with psychobiological development, in which the consensus within the research community is that most traits, even universal and species-typical, use environmental resources non-additively in a continuous interaction between the organism and its environment (see Bateson & Mameli 2007; Griffiths 2009; Michel & Moore 1995). Consider the example perhaps most familiar to philosophers (after Sober 1998a): the development of singing in different bird species. Some species of birds start to sing in a species-typical way[236] even if they have never heard any kind of singing at all. This may be taken as a paradigmatic case of innateness. Other species need to hear other birds of their species to sing in the same way. Merely observing this does not specify the cognitive mechanism of acquisition: whether it involves only a series of triggers or if there is more substantial information processing involved. However, we may consider this as a case of learning. But there are species that are different from both of the above-mentioned types. Members of this third kind do not sing at all unless they are exposed to birdsong, but they begin to sing in their species-specific way regardless of what kind of birdsong they are exposed to. This not a case of learning, nor is it "partly innate, partly learned," since there is nothing in the triggering birdsong that would teach the bird the song it starts to sing. The development of the singing is not plastic or sensitive to the variance in the environmental cue

[236] Notice that species-typicality does not do conceptual work in this example: using species-level traits only has obvious epistemic virtues in talking about developmental processes.

either. The developmental process simply requires a certain kind of external input for certain pathways of development of the singing ability to be triggered.

One proposal could be that innateness is associated with constraints on the development of a trait: development is sensitive to certain factors in the environment, but which and what they do is constrained (for example, Bjorklund and Pellergini 2001; Elman *et al*. 1997). This type of innateness could be called *dispositional innateness*. However, innateness cannot be equated with a mere disposition to appear. Any developmental process usually requires the presence of some specific external enabling conditions: introducing or omitting certain aspects in the environment could actively prevent the trait from developing. This is why the innateness of a trait is often thought to be relative to the "normal range" of environments (see Stich 1975; Sober 1998). But because innateness depends on the causal relation between the trait's development or activation and the environment, the problem arises of where to draw the line between the presence of a "normal" environment and the causal contribution of the environment that would make the trait acquired. In particular, the distinction between what is innate and what is learned becomes blurred: all learning is directed and constrained in some way.

Equating innateness with internal causes or internally determined dispositions does not work, given how biological development works. This has led to attempts to discover a more developmentally realistic interpretation of innateness. A crucial step here is to move the focus from the *causes* of development to its *end-product*. The main function of the distinction between innate and acquired is whether the trait can be affected. I argued for distinguishing the concept of innateness from the folk-biological explanatory context in the previous sub-chapter. An adequate scientific account of innateness can and must be based on a biological understanding of developmental processes that produce a particular end-result robustly and reliably regardless of changes in environmental factors. The attempts in this direction include approaches that I will call *invariance* and *separatist* accounts. The

invariance accounts equate innateness with developmental phenomena in which the end-product is fixed or severely constrained. The general idea was outlined by Steven Stich (1975) and Elliot Sober (1998) and developed further into accounts of innateness as *canalization* (for example, Ariew 1996, 1999 & 2006), *generative entrenchment* (Wimsatt 1986 & 1999), *noninduction* (that is, not being environmentally induced; Birch 2009), and *insensitivity* to a range of environmental variation (O'Neill 2015). The separatists consider the concept of innateness to be a valid concept in fields such as cognitive science and linguistics even if it is not so in biology: some psychological properties can be treated as primitives that a human mind can simply be assumed to possess, without any further explanation being needed, and these assumptions fruitfully guide the research on the field (Samuels 2002, 2004, 2007 & 2009; Kiikeri & Kokkonen 2007; O'Neill 2015; see also Cowie 1999). I will discuss these accounts in the next subchapter. But these accounts have some *prima facie* problems that I will address before getting there.

The problem with separatism is that it does not really give an account of innateness, although it is a strong argument for having one. The problem with invariantism is that all the accounts included in it refer to a specific biological developmental phenomenon that we might call a case of innateness in most cases, but *equating* innateness with either of them may force the reference of "innateness" to move too much for the concept to maintain its semantic closeness to the folk concept. Furthermore, they both raise the question of whether the concept of innateness has a definite meaning at all in the sciences. It is very possible that psychologists and biologists are referring to different things, and as a biological concept, it might bundle together discrete properties. As we have seen, the eliminativists claim that the scientific use of the term "innate" is just an extension from folk biology (see especially Bateson & Mameli 2007 and Griffiths 2002 & 2009). As already discussed, Paul Griffiths (2002) distinguishes between three unrelated dimensions out of which I have narrowed down to Fixity as the core meaning of innateness. Matteo Mameli and Patrick Bateson

(2006) go even further in distinguishing 26 distinct meanings in which the term "innateness" has been or could be used in science[237]. They abandon most of them as vague or nonsensical (and most of the definitions are variants and refinements of others) and they end up with eight "finalists" that they call "i-properties" (ibid: 177–178): "(3) reliably appearing in a particular stage of the life cycle; (12) being such that environmental manipulations capable of producing an alternative

[237] The complete list of their definitions is this: (1) It is not acquired. (2) It is present at birth. (3) It reliably appears during a particular stage of the life cycle. (4) It is genetically determined. (5) It is genetically influenced. (6) It is genetically encoded. (7) Its development does not involve the extraction of information from the environment. (8) It is not environmentally induced. (9) It is not possible to produce an alternative trait by means of environmental manipulations. (10) All environmental manipulations capable of producing an alternative trait are abnormal. (11) All environmental manipulations capable of producing an alternative trait are statistically abnormal. (12) All environmental manipulations capable of producing an alternative trait are evolutionarily abnormal. (13) It is highly heritable. (14) It is not learned. (15) (i) It is psychologically primitive and (ii) it results from normal development. (16) It is not produced by developmental mechanisms adapted to produce different traits in response to different environmental conditions. (17) (i) It is not produced by a mechanism evolved to map different environmental conditions onto different phenotypes and (ii) it results from normal development. (18) (i) It is not produced by a mechanism adapted to map different environmental conditions onto different phenotypes and (ii) it does not result from the impact on development of evolutionarily abnormal environmental factors. (19) It is generatively entrenched in the design of an adaptive feature. (20) It is insensitive to some range of environmental variation. (21) It is developmentally environmentally canalized, i.e. there exists an evolved mechanism adapted to ensure that the development of the trait is robust with respect to some environmental perturbations. (22) It is post-developmentally environmentally canalized, i.e. there exists an evolved mechanism adapted to ensure that the continuance of the trait is robust with respect to some environmental perturbations. (23) It is species-typical. (24) It is a Darwinian adaptation. (25) It is a standard Darwinian adaptation. (26) It is prefunctional. (From a table in Mameli & Bateson 2006: 177.)

trait are evolutionarily abnormal; (18) not produced by a mechanism adapted to map different environmental conditions onto different phenotypes and, at the same time, not produced by the impact of evolutionarily abnormal environmental factors; (19) generatively entrenched; (21) developmentally environmentally canalized; (22) post-developmentally environmentally canalized; (23) species-typical; (25) standard Darwinian adaptation." (The numeration is the numeration of the original candidates; see the footnote above. I will discuss these proposals later.) Mameli's and Bateson's main point is that none of these properties *alone* seems to correlate with the folk notion, and the i-properties themselves do not depend on each other sufficiently to be considered to be properties of the same biological "phenomenon". When biologists use the term, what they are usually really referring to is one of these things. And, as Bateson (1991) notes, if it is something else (and something more precise) that you mean by "innateness", why not say just *that?*

8.2. A Contrastive Invariance Account of Innateness

To sum up, the problems with innateness seem to be the following: (1) Intrinsic versus extrinsic does not seem to be a valid distinction. (2) The folk notion seems to rely on something like this (even if narrowed down to Fixity). (3) The biological properties that could replace "innateness" do not cluster sufficiently. However, there is no need to give up a concept altogether just because it is sometimes used in a confused way or because our folk-biological thinking connects it to an essentialism that scientific biology explicitly refutes. Even if the consequences of some trait being innate were widely misunderstood, there might still be a useful and non-redundant way of using the concept. I will argue for such an account. My account is a variant of the accounts I have labelled "invariance accounts" above. I will start by discussing the earlier invariance accounts, moving to a psychological explanatory context of the notion of innateness and the need to distinguish

between different notions of innateness, then presenting my formulation of the general invariance account of innateness, and finally explicating its relevance for evolutionary explanations.

8.2.1. Invariance Accounts of Innateness

The general idea of innateness as invariance is that the outcome of development is fixed: it is invariant regarding external factors in development. Innateness is contrasted with *plasticity* rather than being acquired – the same genotype produces the same phenotype in all normal environments, even if environmental factors are contributing (but only enabling) factors in development. According to the invariance theories, the development of a trait is innate across a somehow fixed set of environmental factors. An influential account of innateness put forward by Andrew Ariew (1996, 1999, 2006) that equates innateness with *canalization* is an example of this. Canalization is a form of developmental overdetermination found especially in the early stages of development. For example, when basic tissue types start to develop into specialized tissues, the same developmental process can be triggered by different hormones, even though the hormones have otherwise different effects. The development of organs is canalized as well: the organs develop into morphologically identical forms in different internal environments that nevertheless make a causal contribution to the development. The original state, determined by the genome as a whole and the environment, sets the development into a certain canal, in which the causal interaction between the genes and the environment produces the same pre-determined outcome regardless of the specific causal factors. The end states of the possible canals are determined by the intrinsic properties of the developmental system, but the process itself is an interactive process. (See West-Eberhardt 2003; Moore 2003.) The idea of innateness as canalization is that the same phenomena extend beyond these basic contexts, including to psychological contexts. Equating innateness with canalization also means that the distinction between innate and acquired is not a distinction

between genetic and environmental causes, which may be unintuitive to some (see Moore 2003), but we have already moved from defining innateness with kinds of causes to the fixity of end-results.

Other problems arise if innateness is equated with canalization. For one, canalization is not about a trait appearing regardless of the environment, but about the same trait appearing in causal interaction between *several different sets* of environmental and genetic factors. There is a logical continuum between a trait that is dependent on a particular causal makeup of the environment and a trait that is canalized across all possible environments. More importantly, there is no logical difference in the actual developmental process between these two extremes if the causally necessary characters of the environment are reliably present in all actual environments. This is especially true if we are interested in the consequences for evolutionary explanations. Furthermore, the outcome of any trait is influenced by some possible environmental causes (if only by preventing the trait's development). Consequently, the innateness of a trait is always relative to the environment. This may be counter-intuitive but probably unavoidable (see Sober 1998a; Khalidi 2007; Kiikeri & Kokkonen 2007; Birch 2009; O'Neill 2015). This means that there must be a way to constrain the environmental conditions under which the development needs to be canalized (see Mameli & Bateson 2006; Kiikeri & Kokkonen 2007; Birch 2009; O'Neill 2015) and the same requirement follows for any invariance account.

Another problem with equating innateness with canalization is that canalization is a specific biological phenomenon that has evolved to fix certain crucial developmental outcomes. It seems that there are other ways in which the development of the trait can be robust and its end-product fixed across the relevant environments (see Moore 2005; Griffiths & Machery 2008).[238] Either Ariew's "canalization" must be interpreted as a general property of insensitivity, in which case referring to

[238] Strictly speaking, canalization refers to insensitivity to variation both in environmental and genetic factors. This is why Ariew (1999 & 2007) qualifies that he is focusing on the *environmental* canalization only.

the phenomenon of canalization is misleading, or canalization is only one example of innateness (see Bateson & Mameli 2007; Kiikeri & Kokkonen 2007; Birch 2009; O'Neill 2015). Either way, a simple equation does not do the necessary work. Rather, canalization (in the proper developmental context) is a clear example of a developmental phenomenon that counts as innate and of a mechanistic basis that explains it. The accounts of innateness as a property must be less specific and reference to a specific biological phenomenon is not necessary.

There are, however, other invariance accounts of innateness. Ron Mallon and Jonathan Weinberg (2006) refine the basic idea as *closed process invariance*. Jonathan Birch (2009) defines innateness through the contrast of *not being environmentally induced* and Elizabeth O'Neill (2015) as *insensitivity to environmental variation*. I will return to Birch's and O'Neill's accounts later in greater detail and build my own account as an amalgamation of their views. The invariance, however, comes in degrees. Even canalization comes in degrees in two different ways: how wide a range of environmental factors the trait is canalized across and how fixed the specific outcome is. The change from high environment-dependency to developmental fixation across a wide range of environments is gradual.

Both concentrating on the outcome instead of the process and the relativity to the environment may seem unintuitive from the folk-biological point of view, but this also highlights an important difference between folk biology and scientific biology: the former is a metaphysical project, the latter is not. If we abandon essentialism, the relativity of innateness should not be a surprise or a problem. Too much relativity, however, waters down the idea. We cannot just choose the environments in which a particular trait happens to develop and state that the trait is innate in these environments. Every developable trait would be innate in some environments. The concept should be informative to be useful. The graduality of innateness might tempt us to propose a continuum in which one end, "truly and fully innate", would be developmental fixation across all possible environments, and the fewer possible environments are included, the less innate a

trait is. However, setting all possible environments as the point of comparison is irrelevant for any imaginable use for the concept of innateness once we have discarded the folk-biological metaphysics of innateness. We are interested in a smaller range of possible variation – for example, a range of environments normal to the species. After (and only after) we fix the environmental range (which will include a set of causal factors as background conditions), different causal factors may play different roles in the development with respect to the outcome. Some affect how the trait turns out (they are instructive) and some are just a part of the causal network that produces the trait (enabling factors). I will now turn to the issue of the relevant contexts.

8.2.2. Psychology, Innateness, and Primitivism

The main scientific context for nativism debates is not biology but psychology. It seems that regardless of biology, the distinction between innate and acquired has done some work in this context, given that the debates themselves are substantial debates.[239] Furthermore, nativist evolutionary psychology presupposes its object of study to be innate in the psychological context. It would be natural to think that innateness in psychological contexts is the same thing as innateness in biological contexts. However, although the concept of innateness has a similar *function* in both contexts, it seems to have substantially different uses in biology and psychology, as pointed out before by myself

[239] Probably the most heated debate over innateness has been in linguistics. Linguistics is a pluralistic study of language and there are different ways to understand its object. The most obvious one is to study how language as a public medium of communication works, and what kind of regularities and phenomena can be found in it. Another is to study language as a psychological phenomenon. Whether there are language-related specific cognitive capacities that enable its use and acquisition, and whether the development of these capacities and the acquisition of language using these capacities are innate, is a psychological question.

and Mika Kiikeri (Kiikeri & Kokkonen 2007) and Elizabeth O'Neill (2015). The concept of innateness is used as an explanatory concept in both contexts: it has to do with what guides the development of a trait. But when the context of explanation is shifted to a different level of abstraction in description of the processes, the relevant aspects of the environment change, and the relevant contrast between what is innate and what is not is likewise shifted. The interest might be in the same developmental process, but different things are included in the causal background and different interventions are relevant.

The biological interest in the development of a trait is in the biological processes guiding it, and innateness in biology has to do with independence of the developmental outcome from any relevant causal features. Psychological interest is in the psychological processes and their role in development: if the trait, such as mastery of the syntax of a language, develops in any linguistic environment without anything that could be called learning taking place, it is innate – no matter what exactly causally guides the development at the biological level. It is trivial that the linguistic environment plays a crucial role in language acquisition by the linguistic environment triggering developmental pathways in the development of linguistic cognition, and a linguistic environment with different syntactical features would trigger different syntactic competence, but the idea of innateness is that the acquisition of the syntax of a particular language does not involve *learning* of the syntax, understood as extracting it from the informational environment. The debate on whether "language is innate" is about the *nature* of the interaction between the developing linguistic cognition and the linguistic environment. This is why, for example, Muhammad Ali Khalidi (2002 & 2007) has defended a definition of innateness in psychology that does not reduce psychological innateness to a certain kind of biological developmental process but refers to triggering and poverty of stimulus in an old-fashioned way.

Before returning to the more complicated example of human language acquisition, let me return to birds. The birds that learn the singing are usually not able to learn just any singing whatsoever (although

there are some birds that are very flexible in their learning capacities), so their learning capacity is strongly directed to a certain type of song at least. This, however, is always the case with learning to a certain degree – there is some sort of capacity for acquiring and some capacities are more limited than others. This is one problem in drawing the line between learning and innate developmental processes. A capacity that is directed to learning only one kind of thing might lie on the borderline, but we can always use the example of a bird-species that can learn several different songs. The main idea is that the bird somehow imitates a feature of the environment. The environment does not just trigger the singing, but there is some sort of information processing from the environment. Contrast this to the third type of birds that are a clear case of not being a clear case. They need to hear birdsong to start singing but it does not matter what kind of singing they hear. It does not even have to resemble their species-typical singing, but nevertheless they start singing in a species-typical way. This is usually considered to be a further example of the fuzziness of the concept of innateness, but I would argue that there are two different levels (or foci) of explanation and hence two different concepts of innateness here.

The main idea is that the singing birds of the third kind do not possess biologically innate song, since its development is causally dependent on a specific external factor that cannot be considered part of the causal background for developmental psychobiological explanatory purposes. But it is not learned either. The development does not utilize any psychological mechanism by which the bird acquires its singing from what it hears by processing information. The environment is part of what guides the development and determines the outcome, but in a more brute-causal way than in learning. However, this process utilizes *some* cognitive capacities that are needed for the bird to be exposed to the triggering factors in the first place. There are two possible intuitions about innateness here – one has to do with the development's dependence on a specific causal input, the other with this not being a case of learning.

The biological and psychological levels of description and explanation are interested in different kinds of factors in the third case. From a biological point of view, the environment consists of all the features of the environment that are causally relevant to the development. Some of these are part of the causal background and can be left out of consideration. Biological innateness would be insensitivity of the developmental outcome to the other causally relevant features of the environment. Dependency of the outcome on a specific feature of the environment makes this not a case of innateness. To simplify the situation a bit for now, the relevant environment for psychology is the informational environment: the features of the environment that the bird can perceive as units of information for its cognitive processing. If the bird were to gather information from what it heard, this would be a case of learning. It is not: the outcome of the development is insensitive to the informational details of the singing that the bird hears. The singing only triggers (or participates in) the relevant developmental processes producing the singing. Hence, from the psychological point of view, this is a case of innateness.

There is no clear-cut distinction between learning and highly specialized capacities to react to specific features of the environment, and the distinction between a causal process and information processing is not clear either. Therefore, I do not attempt to distinguish between psychological and biological innateness with the concept of information. For now, I am simply trying to make a point about them being different concepts – or the same concept (playing the same conceptual role) but having a different extension in different contexts, with different substantial conditions. What the concept of innateness is intended to do in the biological context and what it is intended to do in the psychological context are different things. Still, they face similar problems (probably inevitable vagueness in borderline cases), and there is no simple way to draw the line between causal contribution to the psychological development and being an information source for a learning process. But equating the two types of innateness still causes greater confusion. Furthermore, the notion of psychological

innateness is compatible with how the concept is used in the language acquisition debate, and confusing it with a stronger biological notion of innateness is a source for confusion in the debate.

Let us suppose, for the sake of argument, that something like the Chomskyian view of language acquisition (see Chomsky 1980, 1986, 1988, 1993 & 2000) is correct.[240] This means that developing mastery of syntax in the first language involves more than just learning. The environment does not provide the information that the child needs to develop the mastery of that particular syntax just by learning from it ("the poverty of stimulus" argument). However, the acquisition of language is not independent of the triggers of the linguistic environment even in the most radical version of this idea. There is a predisposition to pay attention to certain features of the linguistic environment and the clues in this environment trigger developmental pathways that eventually lead to the mastery of the syntax of the language that the child is exposed to. Since certain features (syntactic characters of this language) causally guide the development to a certain direction (to the mastery of this language), the syntax that the child develops is not biologically innate. It "grows" under the precise causal guidance of the environment. In fact, the specific language determines the syntax that the child learns. But the child does not learn the language from the contingent features of the linguistic environment through "general" learning either, only by making hypotheses and testing them. Instead, there are specific features in the structure of the child's psychological development that predispose them to pay attention to specific features of the overall auditory environment, and there is a limited number of possible developmental pathways that can be triggered by these features. If this model of acquiring the syntactic rules of a language is correct, it is not a case of learning, but it is still controlled by the environment.[241]

---

[240] Noam Chomsy's theory of language acquisition has not remained the same in details but this does not matter for the illustrative purposes.

[241] I am not defending nativism about language here. Chomsky seems to be wrong at least in his characterization of the passivity of the child in language acquisition. He says explicitly that "[l]anguage learning is not really something

The Chomskyian way to conceptualize this is that the child has *innate knowledge* of possible linguistic structures. Whatever this knowledge of possible structures of a language is supposed to be on the cognitive level of description, it is an abstraction of cognitive dispositions that translate into implicit knowledge *during* and *after* the learning process – it is not explicit knowledge of abstract rules prior to that. On a biological level, this process has to do with what neurological developments are possible and what constraints there are on development. Which pathways of development are taken is guided by the relevant aspects of the linguistic environment that interacts with this process. Even if certain properties of the syntax that the child develops mastery of are innate in this sense, their actual development is guided by the environment and depends greatly on certain aspects of the linguistic environment. An alternative way to conceptualize the same phenomenon is to say that there is an *innately primed learning process* (see Pullum & Scholz 2002). I consider these two conceptualizations alternative descriptions of the same process.[242] The innateness of syntax in the Chomskyian sense is psychological innateness, but not biological innateness.

---

that the child does; it is something that happens to the child placed in an appropriate environment" (Chomsky 1988: 134). However, children are active in language use from early stages on, and active participation in communication seems to make a difference between children in how efficiently they learn (Romeo et al 2018). This does not, however, undermine the basic idea that the structure of human psychological development has language-specific features that guide and constrain language learning specifically. If there were none, this would be quite anomalous among developmental processes of complex features of human psychology. See Cowie 1999 and Pullum & Scholz 2002 for the criticism of the poverty of stimulus argument; see also Khalidi 2002 & 2007; Kiikeri & Kokkonen 2007; Dediu 2015; and O'Neill 2015.

[242] The historical context for Chomsky's theory of innateness was to make a contrast to the Skinnerian behaviourist theory of language learning (see Skinner 1957; Chomsky 1959). The accuracy and adequacy of this criticism can be disputed (see Palmer 2006), but Chomsky's main point, in my estimation, has always been that learning of language involves more structure from the cognitive process than general learning mechanisms. Regardless of how

It also seems that a part of the nativism debate on language is about the specificity of langue-related skills, not innateness itself. It is enough for innateness that there are guiding and constraining features in the developmental system. Terrence Deacon (1997) turns the Chomskyian idea of positive guidance of "innate knowledge" about the syntactic structures, appearing as the abstract "Universal Grammar", into the limitations that human psychological capacities place on the possible syntaxes. These limitations, too, can be expressed as an abstraction of the rules of possible grammars – that is, the "Universal Grammar". If we adopt a systems approach to psychological development, these are just two different approaches to the same phenomenon. Deacon's main point is the co-evolution of language and the human brain: the earlier stages of the evolution of language were constrained by more primitive capacities, the importance of language placed selection pressures on our abilities to learn language, and the evolution of the brain enabled further evolution of language. (See also Tomasello 2008 and Donald 1998, 2001 & 2017 for a co-evolutionary view.) As discussed in the previous chapter, there is nothing peculiar, from the evolutionary point of view, in evolved developmental mechanisms where the outcome depends on both the internal features of the developmental process and specific environmental triggers – or in a learning process where paradigmatic learning (like the learning of the meanings of words) and blunter influences from the environment are mixed. On the other hand, even if language significantly influenced the evolution of our cognition (which it most certainly did), this does not necessarily entail language-specific capacities. Yet it would entail innateness in language acquisition. Let me explain this by returning to William Wimsatt's idea of generative entrenchment.

---

successful his theories are about what this involves, this main claim about "innateness" is hardly a radical or even controversial claim within the context of contemporary psychology. Chomsky's more specific theories of language acquisition are an entirely different topic that we do not need to touch on here.

Wimsatt (1999) proposes that "innateness" could be replaced by "generative entrenchment" in most contexts, although he explicitly says that it is not a suitable definition for or a theory of innateness. In both the developmental and the evolutionary process, traits do not evolve independently of each other, but instead depend on already developed and evolved structures, and these dimensions are connected (Wimsatt 1986a, 1999, 2001 & 2007; see also the previous chapter). The "deeper" traits might also develop in interaction with the environment, but their development is more fundamental, more rigidly fixed, more likely to be canalzed, more likely to rely on stabile causal factors in the environment, more likely to be species-typical, and so on (Wimsatt 1999). This applies to psychological contexts of "innateness" as well: development is constrained, and the characteristics of the developmental system and its possible pathways direct how the causally relevant factors of the environment are able to influence development through their interaction (see also Elman *et al* 1996; Bjorklund & Pellergini 2001; Kiikeri & Kokkonen 2007). If we fixed the context to the environments where a child is exposed to a linguistic environment (which is almost always the case), this alone could account for linguistic development having general features that cannot be explained by children always learning the same things in the same ways. The earlier stages that are generative for further development (in Wimsatt's sense) explain the apparent "generative grammar" (in Chomsky's sense; Chomsky 1986), which should be understood as an abstraction of the systemic features of human psychological development. This depends, of course, on the theory of learning used. And, once again, I am not defending this view about language acquisition as such. However, if language acquisition involved something like this, some of its features could be characterized as psychologically innate (they appear regardless of the details of the environment) but not biologically (there are specific features in the environment that trigger the neural developmental pathways). And more generally, generative entrenchment is a common property of developmental systems that

may explain some robustness in outcome without a specific (evolved) mechanism that ensures this, such as canalization.

The idea that characterising something as innate has different aims in biology and in psychology has been pointed out before. Fiona Cowie (1999) considers innateness to be a black box for biologists that is opened in psychology.[243] Richard Samuels (2002, 2004, 2007 & 2009), in turn, has defended a version of nativism in cognitive science that he calls *primitivism*. According to him (Samuels 2002: 246), a psychological structure S is primitive if and only if

1) S is a structure posited by some correct scientific psychological theory; and

2) there is no correct scientific psychological theory that explains the acquisition of S.

According to primitivism, psychological primitivity is what is meant by innateness in psychology. It is worth noting that Samuels refers to *correct* psychological theories and explanations. Innateness is not relative to the current status of theorising within psychology (Samuels 2002: 246, fn. 20).

The primitivism account seems to capture the general idea of what I have been characterizing as psychological innateness. It has some limitations, however. The second condition does not apply to cases where abnormal environmental conditions affect the development of brain directly, where the acquisition of a psychological characteristic is not characterizable by a psychological theory, yet it is a psychologically relevant developmental process. Samuels acknowledges this by adding a normalcy condition, but this is not completely satisfactory to cover all relevant cases (see Khalidi 2007). Furthermore, it does not account for innate *differences* between people in

[243] Cowie belonged to an eliminativist camp herself. She did not think the notion has a clear meaning and did not argue for a separation along the disciplinary borders, but considered the psychologists' black-boxing a division-of-labour issue.

psychological characteristics or capacities. The definition does not acknowledge that innateness comes in degrees either. These are, however, problems with the definition. I will next argue for an account of innateness that subsumes Samuel's basic ideas. There will be two main differences. First, my account is *perspectivist* with respect to the meaning of the concept and it theoretically allows several well-defined concepts of innateness. Second, in the same breath, I do not consider the above distinction a *disciplinary* issue. I have used the distinction between "biological" and "psychological" as a placeholder for two different approaches, but the difference that counts is the substantial difference between different causal explanatory questions posed of the developmental process.

8.2.3. A Contrastive Account of Innateness

The general notion of innateness that I propose is the following. "Innateness" is an *explanatory* concept that leaves a developmental process, or parts of it, in a black box. It is not about a specific ontology of the developmental process (that is, requiring a certain kind of cause) but about relevant causal dependencies, or the lack thereof. Development depends on causal factors that can be genetic, systemic or environmental, but none of the environmental factors that contribute to the development are explanatorily relevant difference-makers. If the outcome of the development is robust within a relevant *invariance domain* (a range of environmental variation), and this development is insensitive to the factors that are used in the relevant explanatory field, the trait is innate. This makes innateness relative to both a range of environmental variation and the explanatory perspective taken. I will now take a closer look at the accounts of innateness by Jonathan Birch and Elizabeth O'Neill that are close relatives to my account, before going into it in greater detail.

Jonathan Birch (2009) defines innateness as non-induction: a trait is innate if and only if it is not environmentally induced. Environmental

induction refers to a process in which a particular change in the environmental conditions causes a particular change in the outcome (see West-Eberhard 2003; Gilbert 2007; Wolpert *et al* 2010). However, the distinction between induction and mere causal contribution that can change the outcome is notoriously difficult to make. The environmental factor must be *instructive*, not merely *enabling*. One option here is to use the manipulability of the trait with external factors as a criterion (see for example Mameli & Bateson 2006, but Birch considers this too permissive. Instead, he defines induction as follows (Birch 2009: 298):

> *A trait T is environmentally induced in an organism O iff an environmental mechanism features in a causal explanation of why O developed T rather than a trait from a pragmatically determined contrast class of alternatives T*.*

In other words, the inductive environmental factor is the difference-maker within the relevant contrast class. The contrast class is determined by both the developmental dispositions of the developmental system (which traits are possible at all) and the pragmatics of explanation: only some alternatives are relevant. Innateness is the complement:

> *A trait T is innate to an organism O iff it is not environmentally induced.* (ibid)

The strength of this definition, compared to canalization, is that it defines the fixity of the outcome as a general property of a developmental process instead of identifying innateness with a particular developmental phenomenon or a mechanism. Furthermore, it treats innateness as an explanatory instead of an ontological concept, which is also the view argued for here. It is still vulnerable to one of the main criticisms posed by Matteo Mameli and Patrick Bateson (2006; see also Bateson 1991 and Bateson & Mameli 2007) against the very notion of innateness: what is the point of talking about innateness in the first place? What does it

accomplish to say something is innate, especially if it is supposed to be an explanatory concept? Birch's reply to this is basically that calling something innate is shorthand for an unnecessarily long explanation. For example, if we tell a four-years-old that it is innate for humans to develop eyesight (instead of remaining blind), we are giving a sketchy explanation, but an explanation nevertheless (Birch 2009: 299).

This reply is highly unsatisfactory if it is precisely the development of eyesight that we are interested in. As previously pointed out by many, since the very object of research in developmental biology is to understand *how* the developmental process of a given trait unfolds, the very notion of innateness does not make any sense if the whole point is to black-box this process (Lehrman 1953; Kiikeri & Kokkonen 2007; Griffiths 2009).[244] Black-boxing might be good enough for some other fields that are not interested in development, but it still matters whether a trait simply appears (under some pragmatic conditions) or if its existence must be separately verified or explained. This may be the case in psychological sciences, as discussed above, possibly excluding developmental psychology, but also in some fields of biology.

Elizabeth O'Neill (2015) defends another variant of the invariance account[245] which she calls the *insensitivity* account. Unlike Ariew and Birch, she discusses the need for a notion of innateness in psychology specifically, although, crucially, not only in *human* psychology, but also regarding animal cognition. Significantly, she also *relativizes* innateness to specific environmental variation. O'Neill defines innateness as follows (O'Neill 2015: 216):

[244] Personal interactions with developmental biologists provide anecdotal evidence for this, too. They seem to be genuinely puzzled about what "innateness" could even mean.

[245] O'Neill herself contrasts her account with invariance accounts on the basis of the differences that, I agree, are significant refinements to the previous accounts. I will, however, categorize her account as an invariance account in my more general distinction between eliminativism, separatism, and invariantism.

A trait T is innate with respect to variations in an environmental factor F within a specified range of F if and only if the appearance of T is insensitive to variations in F within the specified range of F.

A claim about innateness combines a range of variation in a cluster of factors that, according to O'Neill (2015: 217), "should be determined by the interests of the scientist or scientific community." This is a crucial condition in two ways. First, on doing so, innateness is relativized to specific research programmes, in contrast to both discipline-specificity and universal innateness. Second, this moves the question about the environmental variation relative to which the trait is innate from the *end-result* of the research into the *research questions*. The invariance accounts, as well as separatist accounts, generally approach innateness as something we have discovered or want to discover (being a psychological primitive or a known developmental fixation that we can use as a shorthand in explanations). Innateness of a trait is what we want to learn. Instead, O'Neill argues that the main reason to be interested in the concept of innateness is that the assumption of innateness is "a useful tool that helps scientists ascertain whether the appearance of a trait is influenced by particular environmental factors that interest researchers." (O'Neill 2015: 220.) For example, the claim that there is an innate Universal Grammar guided a productive research paradigm in linguistics – whether we should abandon it now or hold onto it. This is an epistemic point, but an important clarification of the function of the concept of innateness.

The accounts by Birch and O'Neill are quite similar and combinable, and the positive insights of the other accounts seem to be accommodable to them. They both build on the idea of the fixity of the trait across various developmental conditions (instead of a source for this fixity) and relativize the truth conditions to a range of environmental factors and, instead of a normality condition of some sort, to pragmatic interests. I will take this general idea, Birch's approach to innateness as an explanatory black box and use of the idea of contrast class in the definition, and O'Neill's relativization of the domain of

interest as the basis for a definition for innateness. I should also add that I have presented, independently of Birch and O'Neill, and jointly with Mika Kiikeri, a notion of innateness that belongs to the same family (Kiikeri & Kokkonen 2007). We defined innateness as fixity of the end-result of the developmental process in which the causally contributing factors of the environment are not included in the explanatory resources of the given domain of research. This results in distinguishing between biological and psychological concepts of innateness, where the first is insensitivity to any biologically relevant causal influences and the latter is insensitivity to psychological explanatory resources, which we understood in terms of information processing (similarly to and independently of Khalidi 2007). We concluded that eliminativism about innateness may be warranted in the biological realm, but innateness might nevertheless be a useful concept in psychological research, while whatever developmentally explains the innateness of a psychologically innate trait will be a developmental biological explanation without a reference to innateness in a biological sense. However, this is not the account presented here. Instead, I define a **contrastive invariance account of innateness** as follows.

> The trait T is innate in organism O with respect to the range of variation of environmental factors F within a research programme R, iff
>
> **Inn 1)** T is fixed across all the possible alternative developmental processes within F;
>
> **Inn 2)** F is the relevant domain of R independently of T; and
>
> **Inn 3)** the differences between $O_1 \ldots O_n$ regarding T are either explainable by non-environmental variation between $O_1 \ldots O_n$ or attributable to abnormal conditions falling outside F (either by range or by factors not included).

In other words, there is an invariance domain with respect to the innateness of the trait, but this domain is not determined by the choice

of the trait (which would make the claims about innateness empty and post-hoc) but by whatever is considered the proper domain of the research field or a research programme: only some contrast classes and counterfactual scenarios are relevant. The claim about innateness is a substantial assumption about fixity within this domain. Not all environmental factors or ranges are relevant to all fields or research programmes. For instance, the context of language acquisition relevant to a linguistic model that maps the structure of ordinary language acquisition and the importance of the informational content of the linguistic environment may include cognitive aspects only and exclude contexts of language deprivation, whereas how auditory triggers specific to spoken language trigger neural pathways may be important to some other explanatory questions. The contrast class of the latter includes the situations in which language is not learned at all, whereas the first does not. Extraordinary end-results (such as an inability to develop language in the first place) and the difference-makers explaining them are not included in the contrast classes relevant to the explanations within R but are changes in the (implicit) background assumptions; in the case of language acquisition, these might include things such as genetic mutations, developmental disorders and abnormally depraved linguistic environments (for example, feral children).

Generally, some psychological models are interested in the "normal range" of psychology (defined by the field) and others in abnormal psychology, and some theories of psychological development are interested in the development of "normal" psychology and others in the causes of abnormal developments. As discussed in the previous chapter, in the context of development in general, regular development may take place in generalizable stages where an appearance of something may not have a specific explanation, while explaining deviations from this does. Deviation may require moving into a different research field and its explanatory resources. Furthermore, the explanation of a deviation may involve an *omission* of something that is usually present (and hence a part of the causal background) rather than a reference to an abnormal positive cause. (See Berk 2013; Narvaez *et al* 2012.)

Shifting into another field of research for explanation may be needed either to deepen the explanation (for example, when a detailed explanation is given for a fixed trait) or to explain an abnormality. However, this does not necessarily involve crossing *disciplinary* borders. Both the models of normally functioning human cognition and the development of abnormal psychology are within the psychological discipline, but they are different sub-fields that have different explanatory interests, aims and resources. Similarly, the various fields within biology may have different explanatory interests. The criticism of the confused use of the notion of innateness put forward by Mameli and Bateson (2006; Bateson 1991; Bateson & Mameli 2007) is probably right in its substance, but different uses of the concept within various fields of biology may also be justified. I will not go further into this issue here. However, I reject the distinction between informational and causal environments as a relevant marker for innateness (contra Kiikeri & Kokkonen 2007 and Khalidi 2007) for a similar reason. Developmental psychology is interested in a plurality of contributing factors in development (see Berk 2013). For example, Darcia Narvaez's theory, which maps various causal factors that make a difference in the development of empathy, includes factors as different as physical touch in infancy, free play in nature, and surrounding adults' attitudes towards strangers (Narvaez 2014; Narvaez *et al* 2016). The overall developmental process is long and complicated, and which factors it makes sense to highlight depend on the specific explanatory interests (what is explained in contrast to what), the structure of the process, and the variability in the relevant factors within the relevant range of environments. The range in turn depends, among other things, on how specific are the conditions that we are interested in, versus how theoretical and all-encompassing our interests are. There is no point in restricting the *types* of causal factors allowed – the aims of the research programme should entail this. Conversely, when the research programme is fixed, various characteristics in the developmental processes or in their outcomes can be fixed as "innate", and the justification for this is internal to the research programme.

The causal background is fixed by the explanatory interests. There is a set of possible counterfactual causal interventions that define the contrast class for the given explanatory resources. A trait is innate when its developmental outcome is immune to this set of interventions. If the explanatory process includes the effects of the constantly present causal factors of the environment, there is no point in talking about innateness in the first place. If, however, we are interested in the behaviour of an animal, for example, it might be useful to keep certain dispositions and capacities fixed and explain others via the influence of the environment. The criteria for the range of environments relevant to innateness are objects for pragmatic consideration. I think this is unavoidable after the revision of the concept from a metaphysical to a naturalistic (explanatory) concept sketched above. But this does not make the concept of innateness arbitrary unless the explanatory projects choose their targets randomly.

One last qualification needs to be made. For any trait, there may be parts or aspects of it that are innate and other parts and aspects that are not. As discussed above, the possibilities of development may simply be constrained by the structure of the developmental system and the end-result determined by environmental factors, but the range can still be said to be innate. If there are general features that can be abstracted from all possible human syntaxes and described as an abstract Universal Grammar, this is an innate part of human language that could play an explanatory role in describing language acquisition, possibly depending on the research programme within which this is done. If the sex of a reptile is induced by the external temperature of the eggs at a certain developmental stage, the potential for male development and female development (and presumably for a range of intersex developments) is innate. An obvious counterargument is that accepting this would make everything trivially innate. But it does not. It means that most traits have some characteristics that are innate, but not everything about these traits are innate, and only within some frameworks.

Now, let us take another look at the i-properties that Mameli and Bateson (2006) propose as the cluster of innateness. As a reminder, they include

i) **reliability** of appearance (at a specific stage of the life cycle);
ii) **evolutionary abnormality** of the environmental manipulations that could produce an alternative;
iii) not being produced by an **adapted mechanism** to map between environmental conditions and the alternative phenotypes (plus the previous condition);
iv) **generative entrenchment**;
v) being **developmentally environmentally canalized**;
vi) being **post-developmentally environmentally canalized**;
vii) being **species-typical**;
viii) being an **adaptation.**

The account developed here identifies innateness only with the first. How are the others related? The i-properties iv–vi are specifications of developmental processes that may *explain* why the trait is innate. Species-typicality does not need to be connected to innateness at all. However, there are reasons why they may be connected *in practice*. First, biology is generally interested in making species-typical generalizations, which means innateness-attributions are also on the species level. Second, if something is species-typical (either universally or in some other sense; see Griffiths 2002), this requires an assumption that there is a species-typical developmental reason for this, and innateness (in the relativized sense) may be a sensible yet fallible hypothesis. Nevertheless, I do not consider species-typicality to be a required entailment for the definition of innateness, and I agree with Mameli and Bateson that using them interchangeably is a confusion. Likewise, there is no reason why references to adaptation of the trait or the developmental mechanism producing it should be conceptually

connected to innateness. Again, these may be connected on the level of substantial biological assumptions. This in turn is the main reason to bring the issue of innateness into the discussion in this dissertation. I will move on to this issue now.

8.3. Nativism and Evolution

Let us return the topic of developmental individualism and holism now and integrate it with the above detour to the innateness debates. There can be natural selection only between traits that are inherited. The contents of individual learning processes (or the resulting behaviour) cannot be selected but the capacity for this learning (based on the resulting behaviour in the selection environments) can, as I have discussed earlier. With more complicated traits, such as language acquisition in the hypothetical case that it involves innateness, learning, and all the processes in between, the object of selection becomes more complicated, too. Language, and many other human traits, including many forms of social interaction, involve social learning and culture and other externalized channels of inheritance that participate in the replication of the trait, which enables natural selection to operate but makes the replication holistic. This means that the evolution of the trait must be explained as a trait of the group sharing the developmental resources – I will explicate what this means in the next chapter. However, the existence of holistic replication does not mean that all explanatorily interesting aspects of these traits or practices they participate in are holistic in this sense, or that all human psychology or behaviour are holistic. We need a methodologically sound way to distinguish between when holistic approach is needed and when development can be approached as an individualist process. Furthermore, we need to know what it means for a psychological or behavioural trait to be individualist in this sense, if the context includes all the factors that enable holistic replication.

My suggestion is that developmental individualism can be equated with innateness in the sense that I have defined it. If the contents of individual learning are not an object of evolutionary explanation, and the holistic aspects of replication are excluded, we are left with those traits (or aspects of traits) that develop with insensitivity to the relevant environmental variation. My other suggestion is that choosing nativism as a perspective of evolutionary study is a methodological choice of what to focus on. Choices like that are always choices to narrow the object of study and give partial knowledge only. We probably need both nativist and holist approaches to understand human psychology and behaviour. Doing so, it is important to understand the limitations of each approach and their division of labour, and not to overstate what a particular approach says about human behaviour, or whatever object of research. As discussed before, this is not trivial when it comes to human behaviour. I have argued for the distinction between psychological, agentive, and behavioural earlier in the dissertation, which helps making some of the analytical work. What I will do now is to explicate the possible extent of nativistic approaches and how this juxtaposes approaches that acknowledge the holistic forms of replication.

8.3.1. What is Innate in Evolutionary Psychology?

Much of evolutionary psychology is nativist: the object of study is defined as the innate structure of mind. This innate structure is also thought to be species-typical and evolutionarily functional. Not all forms of evolutionary psychology make all these assumptions. Furthermore, it does not follow from making these assumptions about the specific research object of evolutionary psychology that mind is entirely about these structures only. This would be an overstatement of the results of the research, no matter how well based they were. For now, I will articulate what nativism in evolutionary psychology means in a charitable interpretation, after which I will move into discussing how this narrows down the theoretically possible objects of such research.

I summarized the assumptions of nativist evolutionary psychology to include the four following theses in the chapter 4.2.1.:

A) The mind has a **modular**, functional structure;
B) This structure is **innate**;
C) The modules are **domain-specific adaptations**; and
D) The adaptations are for the needs of the ***Environment of Evolutionary Adaptedness*** **(EEA)**

Nativism of some sort seems to be a natural assumption for evolved structure of mind – and a potential problem for the whole research programme if it is false. There is a vast literature criticizing nativism (see, for example, Karmiloff-Smith 1992 & 2006; Elman *et al* 1997; Cowie 1999; Buller 2005; Ylikoski & Kokkonen 2009; Smith 2020). Much of this criticism relies on a strong notion of innateness, however – nativism cannot work because developmental processes are not predetermined enough. Subrena Smith (2020) has recently given a formulation for the core problem of nativism: the *matching problem.* She argues that nativist evolutionary psychology depends on the idea that the similarities in the cognitive architecture between the ancestral humans and contemporary humans must be connected in a "homology-like" link by being hard-wired. If there is no such link across the lineage, the past does not explain the present. Even if we knew that the cognitive architecture and behaviour were similar, we would not know if this was because of hard-wiring or because of developmental reasons: similar experiences produce similar architectures. Furthermore, the architecture may be different because of plasticity of mind, and whatever stays the same down the lineage cannot be as detailed as what the nativist evolutionary psychologists think. The core methodology of matching contemporary cognitive architecture with ancestral cognitive architecture is not justifiable.

The matching problem does not need to be crucial, however. The level of detail that the nativists seem to assume the innate mind has is undeniably problematic and I do not defend it here. However, as I

have already argued in chapter 4, it is not crucial *how* finely-grained the structure is, since this only affects what level of specificity the object of evolutionary psychology can have. Similarly, the mind could also have a fine-grained modular structure that is not innate, but a product of environment-sensitive individual development (see Karmiloff-Smith 1992, 1998 & 2006; Buller 2005; Smith 2020), and in this case the *explananda* of evolutionary psychology would not be the modules but something more general. But the crucial issue is the innateness. How do we interpret innateness in this context and how big of a problem is the matching problem?

As discussed in the previous chapter, natural selection has two targets: the outcome of the developmental process and the developmental process itself. For the latter, reliability of production of the selected outcome is the key issue. Therefore: "innateness." It does not matter, however, what causal factors are included in the developmental interaction. Even if natural selection worked only on the variation in the genes, the outcome for which the genes are selected could still be partly produced by external environmental factors, as long as the specific factors are reliably present in all the developmental environments in the EEA or the development is canalized across the variation within the EEA. If a particular feature of the environment is reliably present in practically all environments where the members of a species grow up, like the singing of fellow members of a bird species, it is just as good as if the singing were encoded into genes. There is a difference between a developmental process that involves learning in the sense of imitation or other cognitive processes, and a developmental process that only utilizes the auditory environment in some more direct causal way. The learning route may be selected against because of the cognitive burden it poses, but dependence on a specific environmental factor that is reliably present may not. This is an evolutionary reason for the fact that we find "strange" cases where, for apparently no reason, the developmental process relies on a seemingly random external developmental resource.

The same applies to human psychology. If there is selection for a specific psychological trait, its development must be passed on to the next generation reliably. This does not mean that the influence of external factors is minimized. In other words, if we make the crude distinction between biological and psychological innateness, where biological innateness is insensitivity to difference-makers that are relevant to biological explanations and psychological innateness is insensitivity to difference makers that are relevant to ordinary psychological explanations, what matters for evolutionary psychology is *psychological* innateness, not biological. Evolutionary psychologists also seem to explain phenomena within the framework of psychology. It is both charitable and probably more correct to assume that evolutionary psychologists make assumptions of innateness in this sense and not in any more demanding sense, upon which the criticism is usually based. However, even if there are psychologically innate evolved traits that are not biologically innate, it means that the opposite of being (biologically) innate is not being psychologically modifiable. If there is no individual or social learning or cultural transmission with regard to these structures, and the developmental interactions produce the same outcome in all the relevant environments, this is sufficient for the nativist methodology in evolutionary psychology.

Things are more complicated than this, however. If the research programme and its aims define what claiming something to be innate means, and the explanatory interests and the relevant contrasts determine the kinds of difference-makers that are relevant, the straightforward distinction between psychological and biological does not apply always. Biopsychology, for example, is not restricted to the same explanatory types as psychology proper. Furthermore, if the developmental environment has changed, the psychological development may also be different with respect to the psychological structures. A new environment could change even an ancient developmental process that is canalized across a wide range of other environments, but the relatively new adaptations (in the evolutionary scale of time) that

evolutionary psychologists are mostly interested in have not had time to adapt to a wide range of environments.

One lesson to be drawn from this is that evolutionary psychologists should draw more attention to developmental interactions with the environment. And there are many examples of this.[246] But another lesson is to be more precise and less ambitious in what exactly is the *explanandum*. I have discussed the difference between a trait as a psychological trait (capacity or inclination) and a trait as a form of behaviour in some of the previous chapters. All the behaviour that is associated with a given psychological trait is what is selected for or against, and all the psychological traits that are associated with a form of behaviour that is adaptive are selected for. There is no one-to-one mapping. Now we must expand this with a developmental dimension. Take fear as an example. If there is an innate capacity to develop fears of certain things, but what exactly an individual ends up fearing is a matter of personal experiences and observing others being afraid, then it is the capacity to develop fears that is being selected, not the individual fears. The usefulness of learning to fear the specific things that people learned to fear within the EEA is the evolutionary explanation for why people fear the things now, but the proper object of the evolutionary explanation is the developmental capacity to develop fears. It might even be the case that no fears were triggered during the individual development, in which case the person would feel no fear. Still, in case they developed fears, fear itself as a category would be innate, and the *explanandum* of evolutionary psychology, but not any specific fear. This move may seem like a trick: in the end, everything is (partly) innate.

[246] To mention a few examples, Karola Stotz (2014) has outlined a theoretical approach to evolutionary psychology that would take developmental plasticity and non-genetic forms of transmission into consideration; Darcia Narvaez (2014; Narvaez et al 2016) has a multidisciplinary, empirical research programme that pays attention to precisely the evolved developmental processes; Christina Moya and Joseph Heinrich (2016) have proposed coevolutionary psychology as a functionalist perspective on cultural differences in psychology; and the evolution of cultural mind itself is an entire research field of its own (see for example, Schaller et al 2010).

Does this not make innateness empty again? Not if we consider nativism to be a methodological stance and a way to identify the research object instead of a radical substantive statement about human mind.

8.3.2. Methodological Nativism as Methodological Individualism

I argued in chapter 4 that we should approach nativist evolutionary psychology as the first approximation of what it would be to do evolutionary psychology that goes beyond the evolutionary history of mind and attempts to develop heuristic tools to understand human cognition and behaviour. I also suggested in the Introduction that the success of evolutionary psychology should be used in the evaluation of evolutionary psychology as a research programme, not theoretical arguments alone. Theoretical arguments are, of course, part of this evaluation, but only a part. What this means regarding nativism specifically could be something like this. Nativism is a working hypothesis: it is a research programme-related practice to leave some developmental processes in black boxes to answer some other questions (that is, function). We do not need to interpret this to mean something else on a completely different field, and the justification of the nativist assumption comes partly from the productivity of this assumption. As a working hypothesis about mind, nativism is also a way to define the research object.

Some of the claims that evolutionary psychologists make are about very specific capacities (such as the cheater detection module described in chapter 4). This might work occasionally. In some other cases, it may turn out that the development of the capacity under study is plastic or depends on some environmental factors that are not the same as within the EEA. This means that the phenotypical trait does not have a direct evolutionary psychological explanation. The range of phenotypic possibilities in plastic development and the fact that there is plasticity in the development may, however, be an object of evolutionary psychological explanation. In other words, the *explananda* may be things like the capacity to learn certain things in

certain situations. What is innate, then, is the capacity to learn and its constraints. This may sound like an *ad hoc* defence of nativism – there was *something* there after all – and watering down innateness, but it is in line with the concept of innateness that has been discussed in this chapter. Whatever can be considered innate within the explanatory framework is the criterion for what can be an *explanandum* within that field. If a proposed object of nativist evolutionary psychological explanation is not innate, it falls out of the domain of nativist evolutionary psychology, and the focus of research needs to change.

I am not proposing that all psychology ends up being innate in some sense, or that evolutionary psychology captures what is essential about it – I am proposing that nativist evolutionary psychology makes substantial innateness claims about its *explananda* in the very act of taking them as its *explanandum*. Consequently, some of the *claims* of evolutionary psychology (such as tight domain-specificity) may be watered down in the process, and nativist evolutionary psychology may not turn out to be a very useful approach to understanding human mind in general. These are entirely empirical matters about human mind. My point here is only that the assumption of innateness is, or should be, a methodological principle to identify the proper objects of explanation as well as a substantial claim about a specific part of the mind's architecture, not about mind in general. How far this methodology goes in our attempts to understand human mind remains to be seen from the results it can produce and how well these results can be integrated with the rest of psychology and other human sciences. Evolutionary psychology may turn out to be a valid research programme, but by saying far less than it promises to say about humans.[247] Evaluating this, however, falls outside my discussion here.

---

[247] Evolutionary psychology is sometimes associated with the idea of *human nature*. Human nature is an essentialist concept that combines innateness, species-typicality, and normativity. Maria Kronfeldener (2019) has recently presented a revisionist, post-essentialist, and pluralist account of human nature that naturalizes the concept. She distinguishes between three different epistemic roles that the concept of human nature is supposed to have: descriptive,

Although nativist evolutionary psychologists are probably in the camp of Modern Synthesis,[248] I have approached the notion of innateness and its connection with evolution from the framework of the Extended Synthesis in a way that I hope is compatible even with the Developmental Systems Theory.[249] Individual development may rely on the environment, but only some of it is relevant to selection: only those factors that can transfer selected effects. Nativist evolutionary psychology does not need to deny the existence of environmental contribution to innate traits if innateness is properly understood. Furthermore, if a trait involves plasticity, the object of the explanation is the plasticity and its range. In learned behaviour specifically, the capacity to learn, the context of learning, the potential range of learned behaviour, and so on, are the trait. A capacity to learn something is only selected for the resulting behaviour and how it increases the adaptedness of the organisms with that capacity, given that the variety of resulting behavioural traits is better adaptively matched to the variety

---

explanatory, and classificatory. Her definition of *descriptive* human nature is constituted by all the traits that "are conserved over evolutionary time by biological rather than cultural inheritance" (Kronfeldner 2019: 165). If human nature is defined in this way, evolutionary psychology may indeed be in search of human nature. However, Kronfeldner's conception of descriptive human nature is very thin compared to how the concept of human nature is usually understood. What I am saying about the "innate mind" of the nativist evolutionary psychology is something similar and the issues are directly connected. There is a human nature but there is not much of it – and if nativist evolutionary theory studies it, it leaves much of human mind outside.

[248] I am not aware of any explicit endorsement of the Extended Synthesis in this camp and the basic theoretical tools are rooted in the gene's point of view (see Barkow et al 2002; Buss 2014).

[249] There is a strong personal overlap between the critics of innateness and the proponents of the DST (most notably, of course, Paul Griffiths) and naïve conceptions of innateness are certainly incompatible with the DST approach to development. There are no incompatibilities between DST and the invariance accounts, however, as I hope my discussion has implicitly shown.

of the environmental conditions than if there were one fixed trait across all different environmental conditions.

Nativism may be a valid methodological choice when it comes to some psychological structures in isolation, but not when it comes to human behaviour, especially social behaviour. Behaviour is a part of the developing phenotype, and understanding its evolution requires understanding all the factors that systematically affect it. This includes both proximate and developmental considerations. I discussed the difference between psychological and behavioural traits in chapter 5 and I further distinguished between individualistically and holistically individuated behavioural traits. Nativist evolutionary psychologists are interested in behaviour, too, not just cognition, but they link behaviour to cognition directly as part of the function of specific capacities, and in doing so they are restricted to the individualistic perspective in the proximate dimension. I argued that the function of some individual behavioural dispositions cannot be understood except in the light of the role they play in a holistic view of social interactions where the forms of interaction, or interactive traits, compete. This as such could also be integrated into evolutionary psychology. There is, however, the consequence that the evolutionary function of the cognitive characteristic is not what the individual achieves, but what the interaction with others achieves. The perspective on evolutionary functionality must be holistic. In other words, proximate dimension individualism does not follow from developmental individualism – and neither does evolutionary individualism.

As argued in the previous chapter, various kinds of social interaction and forms of culture can make the replication of social behavioural traits a holistic process, either in the interactionist sense (limited horizontal transmission within the group) or collectivist (the whole group acquires the trait). This means that even if the interest of evolutionary psychology is in those features of mind that stay fixed, understanding their evolutionary function may require understanding what their function is in a culturally changing context, where the ancient cultural settings were not only the selection environment for this psychology, but

the psychological characteristics were functional only as they participated in the behavioural traits that are partly cultural. Furthermore, the development of the innate parts of psychology may still be functionally linked to holistic replication systems. Consider language again. Even if certain parts of language cognition were innate in the Chomsky style, the full story of language development (even in this respect) requires the language community that collectively provides the environment that triggers the developmental pathways to the correct syntactic competence. The evolution of language is a complicated issue (see Gibson & Tellerman 2011; Hauser *et al* 2014) but understanding it will probably require both psychological-level and language community level accounts of the various social functions of language.

I have now given the most charitable reading of individualistic assumption of development within the context of evolution of human social behaviour that I am able to. It seems that just like in the proximate dimension, where individualist psychological traits may be connected to holistic behavioural traits, innate psychology may be connected to holistic replication of behaviour. In the proximate dimension, the holistic trait defines what the behaviour is for, which gives the psychology its adaptive function. In the developmental dimension, both vertical and horizontal channels of replication are relevant for understanding what is being selected. In both cases, individuals are evolutionarily bound together, although in different ways. In both cases, behavioural outcome is the primary concern for understanding evolutionary functionality – psychology is only a component part. It is time to bring these topics together now and to move to evolutionary explanation and see what consequences the discussion thus far has for the issue of group selection.

9. Group Selection and Holistic Adaptation

A key goal of this dissertation is to distinguish between methodological individualism and the various plausible forms of holism in the evolutionary explanations of human behaviour. The purely individualist approach takes the adapting units to be individuals (in a social selective environment) whereas the holistic alternatives take human groups to have been the adapting units, one way or another. In other words, if we want to understand the evolution of a social behavioural trait, we should understand how it is adaptive as a part of how the group functions. As discussed in earlier parts of the book, this is a methodological question about adequate causal explanations, not an ontological issue about causal processes in evolution. Even if the causal processes can be traced to concrete interactions between individuals, other individuals, and the environments, and to individual reproduction, the evolutionary explanation (by selection or otherwise) abstracts away from this to more general features and structures of these processes that express what is causally relevant. These structures, however, can make a difference – they can be difference-makers that are relevant to explanation.

As discussed in chapter 3, processes of natural selection can be understood as mechanistic processes through a causal structure that instantiates the mechanism of natural selection. Assuming such a mechanism for explanatory purposes involves assumptions about what connects the behaviour of an individual to the fitness effects in the proximate dimension and how the behavioural disposition is transmitted to the next generation. Both may involve other individuals in a difference-making way, as discussed in the past few chapters. These are two ways in which the group may become a relevant unit for adaptation by making two different kinds of causal connections between individuals for evolutionary purposes. The third causal dimension is the evolutionary dimension itself: the individuals directly increase each other's fitness. This is the case with a group of evolutionary altruists, for example.

In this final substantial chapter of the book, I will look at the issue of group selection – or, more precisely, how the discussions so far are related to it. I will start with a brief look at the levels of selection controversy and how levels, units, and groups can be understood in the first place. I will then distinguish between the various topics of interest in the groups along the lines of the different causal dimensions that I have discussed. I will argue that both proximate and developmental considerations may be reasons to consider some traits to be holistic adaptations, understood as adaptive products of the selection process even if individual selection is the driving force, but they may also constitute a causal foundation for group-level selection.

9.1. The Levels of Selection

The problem of the levels of selection – that is, what levels of biological organization selection takes place on – has been one of the "big problems" in the philosophy of biology and theoretical biology for several decades. I take the hierarchical nature of selection to be a settled issue, at least in theory. Some issues about its nature remain, and I will take a brief look at some of them. There are some derivative issues, such as the nature of biological individuality (for example, Gould 2002; Goodnight 2013a; Clarke 2016)[250] and the major transitions in individuality (Maynard Smith & Szathmary 1997; Sober & Wilson 1998; Okasha 2005a & 2006; Clarke 2014), which are largely irrelevant here, and I will not discuss them. The topic here is levels of selection in a very narrow special case: human groups that consist of individuals interacting with each other. I will now take a brief look at the topic of the

[250] There are many ways to approach individuality in biology (see Bouchard & Huneman 2013) but the one that is relevant for the current issue is evolutionary individuality. This can be understood, following Charles Goodnight (2013), as the *lowest* level of biological organization that is relevant to selection, to avoid presupposing individualism about selection.

levels of selection and articulate the issues that are relevant to this special case.

9.1.1. The Group Selection Controversy and Multilevel Selection

Natural selection is based on competition and this competition has mostly been understood as individualistic ever since Herbert Spencer (1864). However, alternative ideas have been around for just as long. Charles Darwin himself was open to the idea that selection could work on higher levels of biological organization, although he did not study the idea (Darwin 1871; see also Gould 2002). Pyotr Kropotkin (1902) drew attention to the evolutionary benefits of collaboration and mutual aid while living in social groups, based on both theoretical arguments and empirical observations of animals and indigenous peoples. The modern debate about individualism and group selection, however, is usually rooted in the work of Vero Copner Wynne-Edwards and his critics. Wynne-Edwards (1962) argued that population size regulation by individual decrease in reproduction (to correlate population with the food supply at the time) cannot be explained by individualistic selection but requires group-level adaptation that is not directed by individual competition.[251] His critics, most notably George C. Williams (1964) and John Maynard Smith (1964), showed quite convincingly that this idea of selection *for the good of the group* cannot possibly work, and this coincided with William Hamilton's theory of inclusive fitness (Hamilton 1964a & 1964b), which seemed to be able to reduce seemingly group-level selection to individual selection, as discussed in chapter 4.[252]

[251] The phenomenon itself had already been empirically demonstrated in the late 1920s and early 1930s by Royal Chapman (1928), Raymond Pearl (1932) and David MacLagan (1932).

[252] For the more complex history of this episode in the history of biology, see Borello 2010.

It is generally agreed that Wynne-Edwards's good-for-the-group selection does not work, although the jury is still out on whether the phenomena he discusses are cases of group selection (Hamilton & Diamond 2012). The more recent theories of group selection are based on the concept of *multilevel selection* or MLS (see Wilson 1975 & 1989; Sober 1984a; Heisler & Damuth 1987; Damuth & Heisler 1988; Goodnight *et al* 1992; Wilson & Sober 1994; Goodnight & Stevens 1997; Sober & Wilson 1998; Okasha 2006): there can be Darwinian selection processes on different levels of the biological organization at the same time. If populations have structure, there may be emergent differences between the different parts (or groups) of the population; that is, differences that cannot be reduced to mere aggregates of individual differences.[253] If the differences are relevant to survival, some groups are fitter, and selection takes place because of this. There is no selection of individual traits for the good of the group but Darwinian selection between different parts of the population on the same level of biological organization. The selection on different levels may have opposite impacts, however, such as in the case of group selection for altruistic groups and individual selection for selfishness within. Wynne-Edward's theory, however, set the stage for the later controversies. It made the individualist position the default and switched the burden of proof to the proponents of group selection (see Sober & Wilson

[253] Charles Goodnight (2015), however, cautions against placing too much weight on the difference between aggregate and emergent properties, since the individuals in the different groups "experience" traits differently in both cases. Furthermore, the difference between emergent properties and aggregate properties is not clear-cut but comes in degrees (see Wimsatt 1986b, 1997 & 2007), and this may also mean that group-level differences that are relevant to selection may be much more common than usually thought (for this approach to the levels question, see Wimsatt 1980 & Lloyd 1988; see Godfrey-Smith 1992; Sober & Wilson 1994; and Okasha 2006 for discussion and criticism). This has no relevance to the topic at hand for the most part, however, so I will not go more deeply into the issue. I will, however, make a more specific argument on these lines in later in this chapter in the case of interactive social traits.

1998; Leigh 2010a; Hamilton & Diamond 2012), and this could explain some of the remaining scepticism towards group selection approaches and biases in interpreting the evidence given for it. Even when group-level selection is accepted theoretically, special conditions are sometimes required to interpret an empirical finding as an instance of group selection.

There are two components in the issue of group selection: a theoretical question about how adaptation works and an empirical issue about whether group selection actually occurs (Goodnight & Stevens 1997). These two are intricately connected. There seem to be both experimental results demonstrating the effects of higher-level selection (see Goodnight & Stevens 1997) and empirical observations of its existence in the wild (for example, Donohue 2004; Weining *et al* 2007; Formica *et al* 2011; Pruitt & Goodnight 2014; Searcy *et al* 2014). What the critics contest is whether these cases demonstrate adaptation processes at the group level (see Gardner 2013b, 2015a & 2015b; Gardner *et al* 2011; West & Gardner 2013; West *et al* 2007a & 2007b). However, this criticism seems to be conceptual rather than empirical. For instance, Andy Gardner (2015a & 2015b) even acknowledges that there is *selection* at group level (that is, based on population-structural differences), but since the resulting adaptations are on the individual level, there is no group level *adaptation* process. This would require selection for group-level properties that works against individual selection. In other words, even if the group structure makes a difference (and even if you call this group selection), it is not selection for group-level traits and therefore not group adaptation or "true" group selection.

This line of criticism echoes Robert Brandon's *screening-off* criterion for higher level interactors (Brandon 1988). Brandon understands group selection as selection for group level interactors (or group-level interactor properties). Since group-level differences may simply be aggregates of individual differences and the selection takes place on the individual level, the higher-level differences must screen off the lower-level effects on reproduction in order it to be real group

selection.[254] According to this interpretation, selection is individualist, and the local social environment is a selective environment. Group selection would occur only when the groups form superorganisms. However, this is clearly too strong a criterion: it ignores relational properties that are emergent structural properties even if they are not independent of the individual properties (Sober & Wilson 1994; Okasha 2006; Goodnight 2015). Furthermore, there would be group adaptation only with evolutionary altruistic individual traits that are optimized for the good of the group instead, which is an unreasonable requirement for both conceptual reasons (conflating adaptation process on the given level with an adaptationist assumption of the process being able to reach optimal adaption on that level) and empirical reasons (since there is never perfect adaptation on one level if there is selection on different levels). (Cf. Pruitt & Goodnight 2015; Goodnight 2015.)

The main lesson from this interpretation controversy is that the multiple selection processes on different levels of biological hierarchy cannot be interpreted as if they were simply different force vectors. The higher levels of biological organization do not behave in the same way as the individuals for selection purposes. Groups are defined precisely by their being groups of individuals – they are not superorganisms. And if they become superorganisms, there is a transformation in individuality. Multilevel selection is a selection process in which differences on multiple levels of biological hierarchy determine what gets selected. Individuals, at least in the case of human groups and groups like them, play a special role in this: they are functionally

[254] Kim Sterelny, too, criticized the multilevel selection of failing to have genuine group selection in it in the 1990s (see Sterelny 1996a & 1996b). He proposed what he called *broad sense individualism* to be an alternative interpretation for how structural properties affect selection that is still somehow a case of individual selection (see also Kerr & Godfrey-Smith 2002). I will lump this view into the broader category of views with overtly demanding criteria for what counts as group selection. Furthermore, Sterelny has later changed his mind on the issue.

integrated, and this is the lowest selectively relevant level of hierarchy[255], and individuals are the replicators. It is a matter of semantics whether one calls the group level of multilevel selection "real" group selection or not. There is, however, a real difference between group properties that evolve as "design" properties of groups and individual properties that evolve because of group-level selection. I believe that this is the distinction Gardner (2015a) makes when he distinguishes between group selection and group adaptation. Elisabeth Lloyd (1992, 2001 & 2017) discusses this as the distinction between the interactor question and the manifestor-of-adaptation question, which are theoretically different issues about levels of selection. I will return to this shortly. However, the group traits in the stronger sense need to emerge before there can be selection between them in this stronger sense, and this emergence is guided by group-level selection in the sense of the multilevel selection theory (see Sober & Wilson 1998; Michod 1999; Okasha 2005a & 2006). The substantial issue is about the relationship between the properties on different levels: should we approach human social behaviour, for instance, as behavioural adaptations of an individual (that is, individual traits), or as interactions on the group level (holistic traits)?

The issue is closely related to another asymmetry between different levels in the multilevel selection model: how fitness benefits function. MLS can be given two interpretations that John Damuth and Lorraine Heisler (1988) call multilevel selection 1 (**MLS-1**) and multilevel selection 2 (**MLS-2**) (see also Okasha 2006). In the MLS-1 interpretation, the differences between groups are the group's contribution to the fitnesses of the individuals in the group (hence the selection process works through individual fitness differences). In the MLS-2 interpretation, groups as *collectives* have fitness (group fitness) and they have more or fewer offspring, understood as different types of groups, even if the reproduction of groups takes place through individuals.

[255] Excluding meiotic drive, which is a relevant sub-individual level selection process.

This is not mere difference in conceptualization but in types of reproduction. There may be one type of MLS process without the other, although they may also be simultaneous and drive the same adaptations. (Okasha 2005a & 2006.)

MLS-1 describes processes in which the evolving traits may be individual or relational while the group structure matters to what gets selected, although group-level properties may emerge as a result. The MLS-2 approach is required when group-level properties are evolving that have emergent properties depending on more precise organization. This is required in understanding major transitions where the adaptive independence of higher levels is evolving, the fitness in different levels become decoupled, and the lower levels are eventually "de-Darwinized" (see Michod 1997; Michod and Nedelcu 2003; Okasha 2005a; Birch 2020), but also in the evolution of some social behavioural traits such as division of labour (Hamilton & Fewell 2013). Furthermore, if there are cultural differences between groups that are maintained regardless of which part of the individual (psychological) variation exists in which groups, this is a case of MLS-2. In other words, all processes with evolving group properties that are not reducible to individual properties include MLS-2. In contrast, the group selection in MLS-1 is about the group environment's contribution to overall individual fitness in comparison to other individuals, including individuals that are similar in their individual phenotype in other groups. The difference between MLS-1 and MLS-2 as well as the difference between the relevance of group structure to evolving individual properties on one hand and evolving group-level properties on the other are two dimensions in which intuitions about what group selection is become diversified. It also raises the question of what constitutes a group for evolutionary purposes and why all evolving structural properties should be about group selection (see Sterelny 1996a & 1996b; Kerr & Godfrey-Smith 2002; Lion, Jansen & Day 2011). I will address these issues shortly.

The different interpretations of MLS are, in turn, directly connected to another vestige from the Wynne-Edwards era group

selection controversy: the causal interpretation of models with individuals, their direct fitness, and fitness contribution from others. Kin selection was supposed to be an individualist model with the expansion of fitness from direct to inclusive fitness, but as it turned out, modelling evolutionary processes with multilevel selection models and with models using inclusive fitness are mathematically equivalent (see Lion, Jansen & Day 2011; Marshall 2011; Frank 2013; Birch and Okasha 2015; Okasha 2016). What is important for kin selection is not kinship as such, but whether the recipient of the benefits is likely to have the same genetic factors that guide the development of the benefiting behaviour. This can be modelled by both generalized Hamilton's rule (Queller 1992; see also Gardner *et al* 2011) and by Price's rule, which is derived from the general Price's equation (Price 1970) by assuming population structure. Price's equation, in turn, simply describes the covariance between genetic inheritance and phenotypic traits, and it can be applied to multiple levels of biological organization at the same time (see Okasha 2006).[256] The mathematical equivalence, however, does not mean that they model the same causal processes. To be precise, the models may trace the same *outcomes* of the process, but they are too abstract to specify all the relevant causal assumptions. At the same time, the motivation behind the different models is in the background assumptions that make different causal assumptions about the process. Accordingly, some theorists consider

[256] Price's equation is a common way to understand multilevel selection formally. There is, however, an alternative, *contextual analysis*, introduced by Lorraine Heisler and John Damuth (1987), that is favoured by others (for example, Goodnight *et al* 1992; Stevens *et al* 1995; Weinig *et al* 2007; Goodnight 2013b). Contextual analysis considers individual traits, aggregate traits (the group means excluding the focal individual), and emergent traits, which can only be measured in the context of the group. It is not equivalent with the Price's equation approach in all cases (see Okasha 2006; Goodnight 2013b) but this is not relevant to the current purposes. The kin selection approach and contextual analysis remain compatible, just focusing on slightly different measures (see Goodnight 2015; Birch 2020).

kin selection to be a special case of multilevel selection rather than equivalent (for example, Sober & Wilson 1998; see also Nowak *et al* 2010), while some others consider them to refer to different causal processes (for example, Okasha 2016; Birch 2017, 2019 & 2020). I will defend a version of the separation stance later in this chapter. What is important now, however, is to distinguish between the fitness effects and the mechanism responsible of them.

As Elliot Sober and David Sloan Wilson (1998) have famously argued, the selectionist logic of kin selection is that of group selection regardless of kinship. What matters is to match the benefits from (evolutionarily) altruistic behaviour to those individuals who have the same behavioural tendencies. This is the causal connection in the evolutionary dimension alone. Relatedness is a way to build the conditions for this in the developmental dimension. (See also Okasha 2016.) Living in kin-centred groups, as many social animals do, is a way to assure multiple reciprocal connections that can enforce each other, and this makes relatedness special (see also Birch 2017 & 2020). This does not change the fact that the *selectionist* connection is on the group level and has nothing to do with relatedness. Furthermore, sharing genetic (or any other reproductive) resources that are associated with the development of altruistic tendencies is not even necessary for group selection. Multiple ways to develop the same tendencies may be selected under the same selection pressures. Recall the earlier discussion about the evolution of behavioural altruism. If there is group selection for it, even different psychological mechanisms may be selected for achieving the same behaviour, and the selection may be blind to both this and the different developmental pathways to achieve them.

To further summarize this brief review of group selection debates, there are three causally explanatory dimensions that are relevant to group selection: the evolutionary, developmental, and proximate dimensions. The evolutionary dimension is about the logic of adaptive processes and the relevance of population structure to fitness benefit allocation. The developmental dimension is about how the replication processes connect individuals in ways that allocate the

fitness benefits such that they support group selection; that is, whether there are inter-individual dependencies in reproducing developmental factors such as genes but also environmental elements, as discussed in previous chapters. The proximate dimension is about what the evolving traits themselves are – are they individual properties or properties that require multiple individuals, such as group properties? These dimensions should not be conflated, but they are connected. Next, I will take a closer look at this and at the different levels and units of selection, especially Elisabeth Lloyd's work, which I will partly build upon. I will also define the holistic approach in the evolutionary dimension alone and give a first approximation of how the other dimensions are related to it.

9.1.2. Units, Levels, and Individualism and Holism in the Evolutionary Dimension

In chapter 3, I sketched an idealizing mechanistic model of how the causal dimensions in the evolutionary explanations of social behaviour are related. According to the model, a behavioural trait being a historical adaptation requires three kinds of mechanistic assumptions. First, the assumption that a natural selection process guided the evolution of the trait – keeping in mind my comments about adaptationism. This is the evolutionary dimension of the explanation. Second, there is an assumption of a robust (psychological, physiological) basis that produces the behaviour in interaction with the environment, so that we can talk about behavioural *traits* instead of incidental behaviour that might not have an evolutionary explanation at all. Third, there is an assumption that the trait is transferred to the next generation in some way that enables its evolution. All these three dimensions may include assumptions about individuals only, or about their connectedness: via the population structure's effects on fitness allocation, via the replication process depending on shared developmental resources, or via the proximate mechanisms for the behavioural traits

being distributed among multiple individuals. In other words, the evolutionary explanations of the behavioural trait may be individualistic or holistic, as I have defined them, and holism has two forms in all dimensions: collectivist and interactionist. Before going into this in greater detail, I will look at some of the previous conceptual work on the issue of units and levels of selection.

The expressions "levels of selection" and "units of selection" are often defined in relation to each other: the unit of selection is the structural unit (an individual, a group) of the type the tokens of which are the objects of selection, while the level of selection is the level of biological hierarchy on which these entities exist (see Okasha 2006 for example). Others make a distinction between the level of selection as the level on which the selection processes take place and units of selection as the entities that evolve under these conditions, which may simply be individuals even if there is group-level selection, for example (see Lewontin 1970; Brandon 1982 & 1988; Kokkonen 2003). This is a relevant distinction given some of the controversies discussed in the previous section, but the same distinction could be reformulated as a distinction between levels and units of *selection* on one hand and the levels and units of *adaptation* on the other hand, retaining the definitional connection between the level and the unit but making the distinction that Gardner, for instance, takes to be crucial, as discussed above (see also Lloyd 1992, 2001 & 2017). Furthermore, the units of *evolution* should be distinguished from both, as the level of biological organization that is the unit of the evolutionary process: the entire population, or lineage.[257] This is definitional for Darwinian evolutionary processes to take place at all (see Godfrey-Smith 2009), but this is not just about framing what natural selection is. First, other evolutionary forces may be at work on the population level too. Second, there may be quasi-selection processes between species, *species sorting*,

[257] For Richard Dawkins, the reply would be genes (see Dawkins 1976, 1982 & 1984), but we do not need to revisit the ontology of evolution here that was discussed in chapter 8.

where two or more species compete for the same resources and one drives the other(s) into extinction, but the selection processes take place within the population (see Vrba & Gould 1988; Lloyd & Gould 1993; Gould 2002; Lieberman & Vrba 2005). Moreover, as discussed in a previous chapter, replication remains another relevant issue. It has been discussed in the context of levels of selection especially since the rise of kin selection and the gene's eye view (discussed in chapter 4), although *gene selectionism* (for example, Dawkins 1984) is clearly a different topic (see Sterelny & Kitcher 1988; Sober & Wilson 1994 & 1998; Lloyd 2017). As I argued in a previous chapter, the units of replication issue is further divided into the replicator issue and the contributing factors to replication process issue.

Elisabeth Lloyd (1992, 2001 & 2017) identifies four basic questions (which all have sub-questions): the *interactor* question, the *replicator* question, the *beneficiary* question, and the *manifestor-of-adaptation* question. The interactor question is about the levels of biological organization on which the selection processes take place. This entails existence of fitness-relevant differences between the units on that level but not evolving adaptations on that level. The manifestor-of-adaptation level refers to the level of organization at which the adaptations emerge. As Lloyd points out, much of the discussion conflates the notions of adaptation as *product of selection* and *engineering* (see Brandon 1978 and Sober 1984 for the distinction; see also chapter 3); understanding adaptation as a product of selection conflates the two notions of level, while the engineering conception requires there to be design on the given level. I will return to this shortly. This distinction, which I have called the distinction between levels of selection and levels of adaptation, is crucial in clarifying some of the confusions in the debates mentioned above. The replicator question is as previously discussed. The beneficiary question is about the ultimate beneficiary of the selection processes – the answer to which is either the evolving lineage or the genes that survive in the lineage. This is what I referred to as the unit of evolution above. This question is about the ontology of evolution and is not important for the current purposes, but it is important to

keep it separate from the other questions. Omitting this question, there are three remaining questions about the levels of selection, as I have already stated: levels of selection issue in the narrow sense, the level of adaptation, and the replicator issue.

I will now revisit these questions based on the discussions of the earlier chapters. Let me begin with the adaptation issue. I distinguished between three forms of sensible adaptationism in chapter 3: *current use functionalism*, *historical explanatory functionalism*, and *ahistorical explanatory functionalism*. In current use functionalism, the adaptive function is the perspective of analysis only. It is minimally explanatory but of minimal use, although it is the form of descriptive adaptationism in behavioural ecology (see chapter 4). Historical explanatory functionalism assumes that the adaptive function also plays a role in the historical explanation of the trait's origin. This has two criteria: being a product of selection and having a function. Ahistorical explanatory functionalism assumes that the adaptive function of the trait is a useful guiding principle for the functional analysis of the organism and its behaviour. As discussed in greater detail in chapter 3, this is a methodological device for discovery that makes assumptions about natural selection having been a causal factor in the shaping of the trait, but more moderate and different ones than historical explanatory adaptationism.

The question about the level of adaptation depends on the notion of adaptation, and not only the distinction between adaptation as product and adaptation as design. As I argued in chapter 3, analysing traits in evolutionary functionalist framework is an instrumentalist choice and its rationale depends on what it is used for. One possibility is to give the trait's evolutionary history an adaptationist explanation – that is, abstracting the natural selection dimension of the evolutionary process as the focus of explanation. Another possibility is the reverse engineering perspective which, I argued, depends on its fruitfulness while making moderate assumptions about historical adaptation processes. It does, however, involve a proximate dimension mechanistic analysis of the complex causal relations between the

individual's capacities and the environment, including other individuals' function, from the point of view of the causal role played by the given factor to the fitness benefits. This is what I have discussed in the chapters related to the proximate dimension of the evolutionary functionalist explanation. I argued that we should abandon the folk-psychological assumption of individualism – that is, that the agent is always the locus of causing behaviour, which is (typically) conceptualized as intentional action. If we are interested in the evolutionary analysis of either behavioural traits or the psychology underlying them, we should take the functionality of a trait in a wider context as a guideline instead. Since ahistorical explanatory functionalism is essentially an analytical perspective (evolutionary design stance), we can choose an individualist or a holistic perspective. However, not all perspectives will be fruitful. The fruitfulness depends on the proximate dimension mechanisms underlying the behaviour, the fitness consequences of various parts of the mechanisms, and how well the functional analysis maps to the proximate interactions. This is an empirical case-by-case issue. However, I also argued that there is a multi-individual mechanistic basis for at least some social behaviour under an evolutionary functionalist analysis.

Another issue to be clarified further is what exactly counts as a group and what group traits are. There are two ways to think about groups in the human context, as previously discussed: a collective of individuals who form a group that lives together and interacts with each other in various ways to various degrees (such as the groups in which our ancestors lived), and a group as individuals in interaction within such a group, such as a coalition or individuals engaging in an organized collaborative effort. David Sloan Wilson (1975), in introducing his group selection model, defined groups as collections of individuals who interact with each other in fitness-affecting ways. Furthermore, this interaction is related to a particular trait that is being selected – hence his notion of *trait group*. In the original model, the groups are collectives of individuals, but later Wilson and Elliot Sober distinguished between *groups as collectives* and *groups as interaction*

groups (Sober & Wilson 1998; my terminology). Interaction groups can be brief temporary coalitions such as pairs, in which the individuals engage in reciprocal altruism, for example. The main idea is that this liberal notion of a group enables to correctly identify the level of hierarchy where selection takes place. According to Sober and Wilson, the evolution of reciprocal altruism is not individual selection but selection between groups of mutual altruists and groups of selfish individuals, just like in the original trait-group model where the groups were stationary collectives (over the lifetime of individuals).

Some critics of Sober and Wilson (for example, Sterelny 1996a; Maynard Smith 1998) consider this too liberal and require some sort of selected group-level properties such as division of labour in order for the groups to count as groups for evolutionary purposes. This criticism fails on two accounts. First, they are making criteria for group-level adaptation, not for group-level selection (see Okasha 2006 and Lloyd 2017 for more detailed discussion). Second, they assume that only groups as collectives can have structured superindividual properties. As I showed previously, the forms of interaction can be selected traits constituted by individual behaviours that are sensitive to other individuals' behaviour. I will return to this in the next subchapter. Furthermore, the other direction of criticism for group selection takes the possibility of structural properties without group structure as its starting point: there is more to social organization than just groups, whether collectives or interactive groups (for example, Kerr & Godfrey-Smith 2002; Lion, Jansen & Day 2011).[258] The first point is crucial for the logic of selection within the MLS model. The second point has to do with the traits that are being selected. The MLS itself is ambivalent about what constitutes the conditions for there to be the kind of interactions that enable group-level selection to be effective enough to

[258] I will not discuss the second point here. The neighbourhood structure model of Kerr and Godfrey-Smith is a model that can be used to understand evolution in structured populations without groups but the interest here is precisely in the groups.

be important. Connecting structures in the proximate dimension is one: the fitness benefits for the different parties are directly connected in participating in the interaction. Connecting structures in the development is another.

As discussed in the previous section, the kin selection model can be interpreted to be a form of group selection, and the kinship itself is irrelevant. I will return to this in the last section of this chapter, but I take this to be true for now. However, the idea of kinship as a part of an explanation is connected to *another* reproductive level connectedness that is a part of the explanation for why group selection works. In the proximate dimension, an individual's fitness depends directly on others, but in kin selection, the relatedness of the individuals entails that the behaviour that benefits the entire group or others in the group[259] is more likely to promote the fitness of those who have a tendency to participate in similar interactions. The Dawkinsian gene's eye view of evolution places the genes to the fore in the replication part of the evolutionary process. However, if we take the position that individuals are the replicators and the replication itself involves multiple factors that are relevant to evolution, genes do not play a special role. There are other factors that affect the development of behavioural tendencies that in turn promote the transmission of these same developmental factors. In the human case, as discussed, the various forms of cultural transmission are highly significant. From the point of view of the logic of selection, there is no difference between genes and cultural transmission.[260] There are, however, two important

[259] There is a difference between these cases. See the chapter on altruism. Note, however, that selection for behaviour that is good for the group is not always altruistic. Collaboration and division of labour can increase individual fitness directly but also involve group selection.

[260] It should be remembered that something being culturally transmitted does not mean that either what is being transmitted or what effects on behaviour are possible could be just anything. Furthermore, under the co-evolutionary approach to human social behaviour, the routes of cultural transmission are ways of maintaining behavioural tendencies, jointly with genetic

differences between genes and culture. First, cultural properties, such as social norms shared within the group, may be independent of individual properties. Second, cultural properties may be properties of the group: for example, group-specific modes in division of labour, or rules of conduct that bind everyone in the group. This means that culture can directly promote MLS-2 processes.

Returning to the initial proposition that I have argued over the course of this dissertation: there are three dimensions, proximate, developmental, and evolutionary in the narrow sense (which is about selection logic), that are relevant to the methodological question of whether human social behaviour should be approached individualistically or holistically when it comes to evolutionary explanations. That is, if we have an evolutionary functionalist analysis of human social behaviour, there are three dimensions that matter for whether we should take a design stance towards an individual or the social group. The evolutionary dimension is about the levels of selection. A methodological individualist approach along this dimension assumes that the behaviour is "designed" for whatever increases the fitness of the individuals themselves. A holistic approach sees the behaviour as a part of wider network of social interactions in the form of what the behaviour does for the group as a whole that increases the fitness of the group. As I stated in the very beginning, there are two forms of holism in each dimension: interactionist and collectivist. In the selection dimension, this is the distinction between MLS-1 and MLS-2: whether "group fitness" is understood as the fitness effects from the social interactions within the group on the individual or as the fitness of the group itself.

The proximate dimension is about the level of biological hierarchy on which the traits to be explained through their fitness contribution are

---

contribution. Cultural change is faster than genetic, but this should not be overstated in evolutionary functional analysis of human cultural lineages or even the species history. Some cultural forms may be universal to our species. The interest here, however, lies in the variation.

assumed to be. In other words, this is the level of adaptation. The individualist approach takes them to be individual traits, the holistic approach to be traits of a group. Along this dimension, the distinction between interactionist and collectivist holism is about what is the group: interaction groups (such as those discussed by Sober and Wilson) or collectives. What counts as a trait and what counts as a group are two sides of the same coin along this dimension. I will go on to discuss that next. Along the developmental dimension, individualism assumes nativism, whereas holism takes culture and other shared developmental resources into account as a group-level way of transmitting traits.[261] There are also weaker interactionist and stronger collectivist versions along this dimension: the existence of cultural forms of transmission that connect and differentiate individuals within the collective group, and the norms, meanings, and other forms of culture that unify the whole group regarding some behavioural traits.

9.2. The Evolutionary and Other Dimensions

Whether to make an individualist, interactionist or collectivist assumption on any dimension is, to reiterate a central point, a case-by-case issue and depends on the empirical facts. All these dimensions are logically independent from each other but holistic processes in one dimension may contribute causally to the holistic processes along other dimensions; the major transformations are all about collectivisation in all three dimensions, resulting in transformation in individuality, for example. I will next discuss in greater detail how the proximate dimension relates to group selection and move on to the developmental dimension in the one after that. Then we are done.

[261] I will discuss how kin selection fits here in the last subchapter.

9.2.1. The Proximate and Evolutionary Dimensions

I have discussed the proximate dimension of evolutionary explanation in previous chapters solely from the perspective of how to perform an evolutionary functionalist analysis on the proximate mechanisms of behaviour. I argued that behavioural adaptations should be analysed in the full contexts of the behaviour, which might include the behaviour of others. If the benefits of the social behaviour are delivered by interaction where the partner (or multiple partners) actively participate in delivering the benefits, the behavioural adaptation is the interaction. The same point has been made in different ways by some biologists who call these traits *interactive phenotypes* (Moore *et al* 1997; Wolf *et al* 1999; Formica *et al* 2017; Montiglio *et al* 2017). Furthermore, Patrick Forber and Rory Smead (2015) have argued that what matters in the game-theoretical models of social interaction are the *types of interactions*, not the types of individual behaviour as such. When this is combined with my take on the evolutionary functional analysis of social behaviour, it can be understood as a holistic interpretation of the game-theoretical models. I have not connected this to the group selection debate yet. I will do it now. I will first discuss David Sloan Wilson's trait group model, especially in the contexts of reciprocal altruism, and I will argue that trait groups define group traits, and this is what is important. After this I will break down the different traits involved in the interaction and how individual and group selection are intertwined. Finally, I will discuss the connection between holisms in the proximate and evolutionary dimensions.

David Sloan Wilson's 1975 version of group selection is quite straightforward. Suppose we have two alternative phenotypes, one evolutionarily altruistic and the other evolutionarily selfish. Suppose the population has group structure and different groups have different (random) frequencies of altruist and selfish phenotypes. All individuals in altruist-heavy groups do better, and the selfish do better than altruists in all groups. Within some range of variables, there will be more altruists in the next generation than in the previous on the

population level. If the groups reform randomly in the next generation, the group-level selection can maintain altruism. As mentioned above, Sober and Wilson (1998) apply this explanatory principle to reciprocal altruism. Interactive pairs, according to them, are trait groups regarding the traits that are relevant to the interaction. The same individual may participate in numerous such groups over their life. In the usual game-theoretical interpretation, pairs that are constituted of A+A (two altruists), TFT+TFT (two tit-for-tat players), or A+TFT, do better than any combination that includes a selfish player.

I discussed reciprocal altruism in greater detail in earlier chapters as an example of an interactive trait. As a reminder, I argued, first, that we should distinguish between behavioural traits and psychological traits, and second, that the reciprocal interactions should be considered interactive traits. There are thus three different kinds of traits: psychological, individual behavioural, and interactive behavioural. Models of social behaviour, including the TFT-model of reciprocal altruism, model the individual behavioural disposition explicitly. Reciprocal altruism is a tendency to help or cooperate unless there are reasons not to – bad experiences of not being reciprocated by the same partner in the standard model, observation of free-riding in OTFT, some models expanding this to gossiping, and so forth. However, the patterns of behavioural interactions that emerge are the traits that produce the fitness benefits, and the interactions are the evolving traits. This is implicit in the model, but it is the driving cause for selection. If the interactions are beneficial, there will be selection for individual behavioural dispositions to participate in such interactions *and* the selection for these interactions over other interactions (or not interacting at all). There can be selection for an individual's interaction disposition only if there is selection for resulting interactions.

Furthermore, there is selection for whatever underlying psychology instantiates the behavioural disposition. As discussed in an earlier chapter, this is multiply realizable, and it is realistic to assume a plurality of psychological capacities and tendencies that participate jointly in decision-making in social contexts. However, psychological

traits are the primary evolving characteristic on the individual level – the individual behavioural tendencies are a function of the psychological makeup within a context. The psychological traits are selected by virtue of what behaviours they entail in the selective context. The context for the multitude of reciprocity-related psychological capacities and tendencies lies in participating in the interactive behavioural traits. Therefore, as I argued in greater detail earlier, the relevant traits in social interactions for the evolutionary purposes are the evolved psychology and the interactive forms of behaviour – not really the individual behavioural tendencies that are used as proxies in the modelling practices.

What consequences does my argument have for the group selection issue? First, although the psychological and behavioural traits coevolve, they are quite distant from each other. In modelling interactions through behavioural strategies (A, S, TFT), it may appear that the interactions are simply (non-additive) combinations of individual strategies. This is not the case, however. Suppose a behavioural context *C* where there is a choice to interact prosocially or refrain from doing so. "Amos playing TFT with Beatrice in *C*" and "Amos playing TFT with Egon in *C*" are not psychological traits but descriptions of individualized behavioural interactions based on the same psychological tendencies. Suppose Beatrice and Egon have different personalities and tendencies – Beatrice is a behavioural altruist and Egon not so much. The interactions with them turn out to be different, but there are no evolutionary grounds to partition Amos's behavioural traits into "interacting with Beatrice in *C*" and "interacting with Egon in *C*". Instead, there is a psychological basis that is involved in interacting with both Beatrice and Egon, and everyone else, in *C*. This overall disposition is what is selected on the individual level. Its fitness, in turn, depends on the types of interaction in which it participates in the totality of the person's social interactions. The interactions with various individuals, however, are the various *interaction groups*, as Sober and Wilson calls them. The interaction groups, or trait groups, are defined by the interactions that have fitness consequences, and these

interactions constitute the traits of those groups – the *group traits*. There is selection between different types of group traits that are instantiated by different interactions, just as there would be selection between different individual traits instantiated by different individuals. Is this a case of group selection? Yes and no.

Let us start with the "no". If we examine the evolving individual traits, they seem to be evolutionarily selfish and selected through individual selection. The psychological capacity to either participate in or refrain from an interaction in the given context, depending on what type of interaction it will be, is the fittest psychology.[262] The group traits, on the other hand, are selected based on how beneficial they are for the individuals who participate in them. Now, this is not a purely individualist scenario since the level of *adaptation* is the trait group level – the *interaction types* are the selected interaction group traits. But the level of *selection* does not need to be higher than the individual level. Consider the original Wilson model of trait groups with selfish and altruistic types. In any given group, altruists are selected against. The group structure is what is responsible for reproduction of more altruists in the next generation – focusing only on the population-level net effect of different selection processes would be an *averaging fallacy*, bypassing the processes themselves (see Sober & Wilson 1998; Okasha 2006). It is a tautology that what gets selected has the highest net fitness within the MLS-1 framework, but this may mask the contribution of the group structure. But this needs not be the case with social interactions that form the structure of reciprocal altruism (in the evolutionary analysis).

Consider a single group (in the sense of a collective) where all interact with everyone else in the behavioural context *C*. They choose different strategies with different individuals, however. This means that the different individual interactions form different interactive traits. In the Wilson model, the phenotypes are the same across the

[262] As discussed earlier, however, this psychological makeup is likely to include psychologically altruistic components.

groups (altruists are always doing worse in interactions with egoists) and the group structure (differential allocation of different phenotypes in different groups) is the causal difference-maker. This is not the case here. Making the behavioural choice is a part of the evolved psychology, and the individual's ability to choose how to interact with different individuals is a part of the selected behavioural capacity. Furthermore, the psychological makeup is selected for the overall fitness effect it has on the individual, just as all traits are selected for the overall fitness effect they have on the individual: the fitness of the psychological makeup depends on all the interactions that it facilitates with all the individuals the individual interacts with. As stipulated in the setting of the example, there is no hidden group structure that would determine who the individual interacts with – but there are different traits that the individuals construct, and the evolutionary basis for this is in the individual benefit. In other words, this is a case of *individual selection for (interaction) group traits*, not a group selection against individual selection as in the Wilson model.

One could argue that partner choice creates a group structure within the collective of individuals. However, I have argued previously that if the individual has a choice in which course of actions they are taking, choosing to not interact with someone is a part of the *individual* trait, which includes all the possibilities that are on the menu for that individual, and this is the fitness effect that is relevant for the individual selection. Furthermore, I stipulated that everyone is interacting with everyone else in this example. But the point becomes even clearer when considering partner choice as group formation. Some individuals form interactive traits and groups, some do not – there is no selection between existing and non-existing groups or traits, but individuals form groups for their own benefits.[263]

[263] Both the interactive phenotype theory (Moore et al 1997; Wolf et al 1999) and the biological markets approach (Noë & Hammerstein 1994, 1995 & 2016) model emerging social interaction as individual selection that results in adaptation on the level of interactions. This is a weak point for individualism as such, however, since the same reservations that I have presented towards

Now, let us move to the "yes" side of the answer. Although interaction group traits can evolve without group selection, I suspect this to be a theoretical possibility only. The evolution of interactive phenotypes also creates conditions for MLS-1. I stipulated above that there is no selective group structure, which includes the assumption that everyone encounters everyone else equally in behavioural context *C*. That is, everyone interacts with all the same individuals. This is an unrealistic assumption. It is more realistic to assume that there is an interaction group structure within the collective; that is, individuals interact with other individuals in *C* unevenly. This means that different individuals interact with different partners, which entails that some individuals participate in reciprocal interaction groups more often than some other individuals with the same psychological makeup. This means that there is *also* group selection for the very same interactive traits. Furthermore, I stipulated that there is only one collective of individuals. If there is a collective group structure (geographically different groups), there will be differences between collectives regarding interactive traits, creating another layer for group selection to operate: groups with more positive interactive traits do better. My argument is not that we should or even could do without group selection with reciprocal altruism. It is that we should make the distinction between selection for group traits and group selection of those traits. Those who say that reciprocal altruism is a selfish trait (on the individual level) are not wrong, and its evolution can be understood as individual selection in principle. But it does not produce only individual adaptations, it produces group adaptations (for forming interaction groups) and the causal factors making the fitness differences cannot be understood without group level perspective. Furthermore, the emergence of such trait facilitates group selection. Let us have a closer look at these points.

---

other individual-based models as modelling individualistic causal processes only apply.

The first point is about the distinction between group-level selection and group-level adaptation. One consequence of the distinction is that there can be group-level selection without group-level adaptation, such as is the case in the basic Wilson model of the evolution of altruism: there is no selection for group-level properties or structures, but a structural explanation for selection of individual traits that would not be selected without the group structure. The other consequence is that there can be group-level adaptations without group-level selection. Given "group selection" has traditionally been understood to be both of these things at the same time – or, as discussed above, even more about group-level adaptation than about selection – this is an argument for the possibility of one part of "group selection" without another. Furthermore, the main issue here is about whether the evolutionary functional design can be found on the group level or on the individual level only, not about the level of selection processes as such.

Second, the conceptual and theoretical separateness of the level of adaptation and the level of selection does not mean causal separateness. As I said, even if interactive traits do not require interaction group selection, they are likely to create a context for MLS-1 too. There is no reason why there could not be both individual and group selection at the same time while both guide to the same direction. Reciprocal altruism may be an example of this, especially if mechanisms like partner choice create interaction group-structure within the collective group. There is a tendency in the literature to discuss multiple levels of selection only when the different levels select different things. The idea seems to be that group selection is important only when it *opposes* individual selection and becomes necessary for explanation. In this case (and probably many others), both levels select for interactive behavioural traits and the individual psychology related to it.

The main lessons from reciprocal altruism can be generalized. Whenever social interaction forms non-aggregative traits that have non-additive effects on the individuals, there is an individualist selection for psychological tendencies to participate in or avoid these interactions, based on the individual fitness consequence. This may result

in the evolution of interactionist traits that are interaction-group level adaptations. This in turn creates conditions for group selection. Group selection, furthermore, may turn out to be stronger than individual selection. The core idea here is that there are two relevant group-level causal dimensions: evolutionary (selective) and proximate, both of which bind the individuals together in their fitness. Both make the group context relevant for understanding adaptation, and both may lead to group adaptation, but the causal structures are different.

My discussion on the proximate and evolutionary dimensions has been about the interaction groups and MLS-1 so far. Groups as collectives are a relevant level for MLS-1 when belonging to the collective is a decisive factor, which also means that the evolving group traits will be traits of a group as a collective. However, this is not a difference in selection process but in the proximate dimension (the level of trait and the level of group). MLS-2 processes involve groups of one kind outcompeting groups of other kinds by producing more groups of the same kind than groups of other kinds produce of their own kind. In human evolution, MLS-2 processes with collective group selection would have included cases where, for example, the differences between collectives caused groups with one alternative property to go extinct more often, and the groups with another alternative to split into two groups and take over the region. As discussed before, culture has often been considered a medium that can introduce, maintain and develop beliefs, behaviour norms, action types, and skills that are shared by all or many members of the group and make a difference between groups that constitutes a basis for group selection (see for example Boyd & Richerson 1985 & 2005; Wilson 2002; Richerson & Boyd 2005; Bowles & Gintis 2011; Boyd *et al* 2011; Pagel 2012; Richerson & Henrich 2012).

Biologist Mark Pagel (2012) has gone as far as claiming that culture (especially language) made hunter-gatherer type groups *de facto* superorganisms and many aspects of human thinking are group adaptations. David Sloan Wilson (1997 & 2002; Wilson *et al* 2000) has made similar points about group cognition. I will not evaluate these ideas, since I am

only exploring the theoretical space of explanatory possibilities – but some aspects of human culture (that have already been discussed) such as the human tendency to rely on acquired beliefs and behavioural scripts, socially enforced behavioural norms, shared meanings, socially constructed roles and institutions, cumulativeness of cultural change, and so on and so forth, seem like they could unify groups internally and make differences between groups in a relevant way. The relevant psychological capacities may even be adaptations for this. For example, culturally transmitted norms for reciprocity in the given context could enforce the evolution of psychological altruism in that context. This could be called the *Baldwin group effect*.

Culture, however, does not need to *unify* collectives, nor make differences between them either, to be relevant to group selection. As discussed earlier, theories of cultural evolution approach cultural entities as something that spreads from mind to mind (competing with alternative cultural entities) and co-evolutionary theories are interested in the co-evolution of mind and culture and how humans copy each other, who from, and why (see especially Boyd & Richerson 2005 & Richerson & Boyd 2005). There may be various traditions within a collective. Some forms of interaction (that are the traits of interaction groups) may be consequences of innate psychological tendencies (as implicitly assumed in the discussion in this sub-chapter), but some may be learned – and some may be something in between. If a form of interaction is culturally available, it may or may not be acquired, whether this is a matter of acquiring tendencies while growing up, an explicit choice, or directed by other individuals (by enforcement, constituting action types, or some other way) who have assumed these forms of interaction. This means that interactive traits may have a cultural basis, too. This, in turn, may constitute a group selection process, for (genetic) selection of psychological characteristics as well. Culture, however, is not an independent proximate force, but a shorthand for a collection of externalist ways in which people acquire psychological characteristics, and the reactions from other people whose reaction dispositions are likewise acquired in those ways. Culture's role in

group selection, therefore, requires another look at the developmental dimension.

9.2.2. The Developmental and Evolutionary Dimensions

As I discussed in the earlier part of the dissertation, in the context of reciprocal altruism, if there is selection for certain behavioural outcomes (for example, participation in a particular kind of interaction), there is a selection for whatever psychological makeup allows this. It may be multiply-realizable, and the selection may result in pluralism, either on the population level (several alternative traits) or on the individual level (several psychological mechanisms at the same time). There may, however, be differences between the ways in which the behavioural outcome is produced, such as reliability, which causes a secondary selection.[264] The same goes for development, as already discussed. The same outcome can develop in multiple ways, all of which will be selected, but there may be differences in reliability, specificity, generality, and so on. But how development works can also affect the individualism–holism issue. If some of the contributing factors of individual development are such that transferring them to the next generation is a "collaborative" project, this can promote group selection as well.

I have mostly omitted kin selection from the discussion so far, but it becomes important now. I briefly commented on the relationship between group selection and kin selection in the beginning of this chapter. I sided with the position that, from the selection point of view, kin selection is a form of group selection: the group selection describes the logic of the selection. At the same time, I agreed that there is also something else going on in kin selection.

Samir Okasha (2016) has argued that there is a difference between causal processes in the paradigmatic group selection examples

[264] Sober and Wilson (1998), for instance, argue for psychological altruism on these grounds, although their hypothesis is still motivational pluralism for (behavioural) altruism.

and kin selection. Although the models predict the same outcomes, they model different kinds of processes. In group selection (such as the Wilson model), the group mean fitness (which is determined by the local individual fitness values) and the allocation mechanism (which distributes the group fitness effect on individuals) determine the individual fitness values. In kin selection, individual fitness values determine the group mean fitness. The group mean fitness is not a difference maker and there is no allocation mechanism. However, as pointed out by Jonathan Birch (2020), this interpretation seems to make an unwarranted causal connection between individual fitness values and group mean fitness. The fitness of the group necessarily depends on individual fitnesses in a constitutive, not a causal way.[265] Birch (2017 & 2020), instead, looks for the difference-maker in the population structure. Kin selection takes place when the individuals tend to interact with their genetic relatives, while group selection takes place when the population has a well-defined group structure. He introduces two measures: K, which measures the structural property of how differentiated the interactions are between kin and non-kin (running from no difference to kin exclusivity), and G, which measures the robustness of social structures (running from no structure through Godfrey-Smith's neighbourhood structure to Wilson-style group structure). The measures form a two-dimensional *K-G space* in which both measures can be high and low.

[265] Okasha and Birch describe the relation as supervenience, but I think we can use a stronger notion here. The problem Birch rises is not necessarily a crucial problem if we do not interpret the causal relations metaphysically (as Okasha does) but only talk about difference-makers (see chapter 2 for the distinction). The problem is, then, that the characterization of the different causal relations does not tell us anything about *why* there is a difference in the causal relations, which is the whole point of explicating the difference as a difference between causal processes or mechanisms. This is what both Okasha and Birch are after, and this is important in understanding the relationship between the two approaches.

I share the intuition that kin-based biases and population-organization based biases are of causally different kinds.[266] Furthermore, the multilevel selection models (and the Price's rule formalization) and the kin selection models (and the Generalized Hamilton's Rule formalization) seem to be aimed at capturing different mechanisms in their conceptualizations although they trace the same selection results and are mathematically equivalent in their predictions of these outcomes (Okasha 2016). I suggest that the key difference is the difference between the causal dimensions that are connecting and dividing individuals into groups. MLS refers to the fitness effects caused (in the proximate dimension) by social structures that bias the benefits while kin selection refers to fitness effects caused by reproductive connectedness. Both dimensions are involved in the process. Modelling with only fitness effects abstracts away from both. MLS and kin selection models are not so much modelling different processes as approaching the same (or sometimes different) processes from different perspectives. Abstracting to fitness effects and modelling them only masks this.

Okasha's (2016) discussion on how the models differ in their implicit causal assumptions when they are articulated with causal graphs seems to point to this direction too. The problem raised by Birch (2020), that the fitness of the group depends on individual fitnesses in a constitutive, not a causal way, could probably be avoided by adding mechanistic detail to the causal graphs (for example, replacing the causal relations between fitness to causal relations between characteristics relevant for fitness before translating them into fitness effects), and the argument would hold. This would complicate the comparison of the models, though – but the issue is not important here. The two dimensions in Birch's own K-G space approach could also be re-interpret this way: The K-dimension measures how connected the individuals are measured in kinship, and the G-dimension

[266] Furthermore, as I mentioned earlier, kinship makes shared traits cluster, which is an additional virtue for effective social evolution. For example, it resolves the Greenbeard problem (cf. Birch 2017 & 2020).

measures the robustness of the proximate dimension structures. Both are factors that facilitate the same basic logic of group selection. As argued in the previous section, the logic of *selection* is the same in the cases of paradigmatic group selection and kin selection (the trait increases its own positive contribution to fitness on the population level by increasing the fitness of individuals sharing it more often than those who do not), even if the connection between individuals is realized in different parts of the instantiation of a natural selection mechanism.

I would add that, as argued in the previous section, the logic of group selection can be implemented both by the organizational structure (where the group structure alone is responsible for implementing the structure – that is, the causal connection is in the evolutionary dimension alone, such as in the Wilson model) and by connection through group traits. The latter, as I argued, does not necessarily presuppose group selection, since group traits can be products of individual selection, which means that this is not an additional dimension to the model – although, as I also argued, it is plausible to think that the emergence of group traits constitutes the conditions for group selection. Taking this into account, the G-dimension should be branched into the *group structure* and *behavioural trait connection* dimensions, resulting in a three-dimensional K-G-B model.

The models and their relationship are not important here, however. The causal adequacy of a model is not important for all practical or even explanatory purposes (although it may be for some; see Okasha 2016). Furthermore, even if an interaction-group trait evolved through individual selection, since the fitness benefits for individuals participating in it come from participating in a multi-individual activity, it would not do harm to *model* this as MLS-1 case. This would be a smaller loss in detail than modelling an MLS-1 process as kin selection. However, it is important that all these connections (in proximate, developmental, and evolutionary dimension) may contribute to adaptation on group level, making group-level perspective a relevant option in any form of evolutionary functionalism.

Sharing genes is not the only way to make a connection in the developmental dimension either. I sided with the individualist model of replicators in a previous chapter and rejected gene-selectionism in all its forms. Genes as such are not important. They can facilitate the relevant connection in the form of kin selection, but they are not the only type of developmental factor capable to this. In the cycle of individuals developing and reproducing, all difference-making, transferable developmental resources are important. For example, the effects that the behaviour of the parents and wider community in the developmental community have on the development processes, including but not restricted to direct teaching, are external factors that can be difference-making transferable factors and are shared by a group of individuals. For instance, some parental effects may affect all the offspring, and the same tendencies may be copied by them for their parenting, making this an external route of parent-offspring inheritance.[267] These effects may be universal, culture-specific, or family traditions, for example. Then there are belief systems, skills, behaviour scripts, norms, *et cetera*, some of which are acquired by some (as individual preferences from what is available), some of which are group-specific, either for interaction groups (they are acquired as part of an interaction type) or for the whole collective (for example,

[267] If a trait is inherited from parents, this can be genetic or behavioural, but also a complex in which there is a "genetically biased" tendency to behave in a way that has a certain effect on development. I mentioned Darcia Narvaez's theory of the *evolved nest* earlier (Narvaez 2014; Narvaez *et al* 2012, 2014 & 2016). According to this theory, the parents and other members of the community create a selected developmental environment. Their creation of this "nest" is a result of both instinctual and leaned (to use a shorthand dichotomy for a complex continuum of developmental interactions) tendencies, both of which have been transferred from generation to generation for most of our evolutionary history. She describes general developmental conditions and takes a normative stance on a particular nest being optimal for human development, but there could also be selected difference-making differences between different family traditions and cultural groups.

normative behavioural expectations). Their importance is this: cultural factors that modify an individual's behaviour may function in an analogous way to genes when it comes to group selection.

Cultures can enforce behavioural unity or group-wide organized division of labour, making groups act as super-organisms, but this is not a proximate level connectedness, rather a result of a shared source for acquiring the behaviour.[268] There can also be behavioural types that are copied by some, and they are selected if acquiring them promotes their further spread. Theories of cultural representations (Sperber 1996) and memes (Blackmore 1999) that approach the spread of some such types as ideas adapting to human minds as a separate layer of evolution. This layer alone explains only temporary fashions at most, if not amended with a long-term co-evolutionary aspect. Co-evolutionary theories (see especially Boyd & Richerson 2005; Richerson & Boyd 2005; Bowles & Gintis 2011) are interested in how cultural group-level differences affect biological fitness in the long run, causing selection between groups on the basis of cultural lineages and promoting both certain (externally inherited) cultural forms and the evolution of mind to be biased towards those forms (see also Tomasello 1999; Sterelny 2012). What I am suggesting here is that this process can take place both between collectives and between traditions within a collective. In the latter case, *cultural relatedness* through external inheritance (by whichever mechanism) works the same way as genetic relatedness.

The idea of cultural relatedness was introduced by Luigi Luca Cavalli-Sforza and Marcus W. Feldman (1981). Jonathan Birch (2017) has recently developed and defended the idea of cultural relatedness regarding *cultural fitness*. He distinguishes, as I do above, between cultural differences that affect reproduction directly and those that act upon cultural fitness and hypothesizes that they were decoupled in the course of human evolution. This is plausible: being cultural and

[268] The behaviour may be enforced by others, which is a proximate source, but even in this case, the enforcers have acquired the shared norm.

adapting through culture is a key human adaptation, and the cultural evolution may have become too fast for the biological evolution to react even if culture has an impact on biological fitness. Pure cultural evolution is outside my topic, but Birch's idea of cultural relatedness is still helpful. The idea of cultural fitness is the capacity to attract apprentices (see Sterelny 2012). Networks of interaction bias the spread of cultural variants[269] such that the individuals affected by one variant are more likely to interact with each other. This may promote culturally spread norms of prosociality analogically to kin selection – and this may have been one of the driving forces in the biological evolution of prosocial tendencies.

---

[269] Birch defines "cultural variant" as "(i) a property of an individual that (ii) varies between individuals, (iii) originates, at least in part, in a process of social learning, and that (iv) admits of a quantitative characterization." (Birch 2017: 196.) The definition is deliberately vague to abstract away from the concrete processes for modelling purposes (see also Richerson & Boyd 2005; Lewens 2015). Note that I have been even vaguer about what cultural factors are, for opposite reasons: it is a pluralistic category of factors. Co-evolutionary theorists often characterize cultural variants as information acquired in social learning, but this is not the only form in which culture can affect an individual. None of the items in the definition above is a necessary condition for a contributing cultural factor in general. The properties of the given factor determine what role it can play in evolution. Furthermore, talk about "cultural variants" is just as vague as talk about memes (see Ingold 2007; see Sperber 2000 and Boyd & Richerson 2005 for the problems of memes as abstract entities with causal powers). Genes have a robust physical manifestation, but cultural factors are a multitude of different causal connections between individuals. Some cultural factors, however, come in types with connecting causal histories: the same kinds of factors affect several individuals and there are nonarbitrary reason for this. I have been calling this an interactionist form of holism in the developmental dimension. There is sameness in behaviour because of similar, connected developmental processes. The definition is not important here, however. Birch's cultural kin selection model captures the basic dynamic in an abstract theoretical framework even if the idea of a cultural variant remains ambivalent.

Biological and cultural kinship are similar in connecting individuals in the developmental dimension and promoting the evolution of shared traits through the "common fate" of the inherited developmental factors. Furthermore, cultural connectedness clusters properties just as kinship does, which enables the effective evolution of complex behavioural traits. There are also significant differences. Kinship is based on genes, which are a part of the fixed guidance of development, alongside with the non-changing features of the environment. Some of the cultural developmental environment (the behaviour of others that is practically always present, existence of language in the environment, *et cetera*) may be part of the fixed guidance of development for all practical purposes, as discussed in the previous chapter. The space of developmental possibilities that is created by these factors is what we can consider innate psychology. It evolved in the context of culture existing in particular social structures (see Tomasello 1999; Boyd & Silk 2003; Bowles & Gintis 2011; Sterelny 2012; Narvaez 2014; Birch 2017) – but culture and social structures have changed, changing both the context in which evolved psychology operates and what the resulting behaviour is.

Why is all this important to thinking about the methodology of evolutionary human social sciences? After all, even the nativist evolutionary psychologists have always taken something like this to be their starting point.[270] There is, however, a further assumption that the evolved psychology is fixed in its manifestation, and the social behaviour it involves is directly connectable to that in the ancestral environments. Its function within the overall context may be different, but this does not matter for evolutionary social psychology. As I argued in the previous chapter, focusing on the fixed (innate) aspects of human psychology is a possible methodological choice, but it is likely to have serious constraints (even if there were some innate modules, there

[270] For the explication of the fundamental assumptions of the paradigm regarding this, see especially Cosmides & Tooby 1987, 1992, 2005a & 2005b; Tooby & Cosmides 1989 & 1990a; Buss 1995 & 2014.

might also be innateness in the form constraints and directedness in development, which may be articulated in abstract only, similarly to Chomsky's "Universal Grammar"), and the evolution of many psychological capacities may have taken place in a social environment that has also been a part of its developmental environment. Alternative approaches integrating developmental considerations have been proposed, for example, by Kim Sterelny (2003, 2007, 2014) and Karola Stotz (2014), and practised by Darcia Narvaez (2014; see also Narvaez *et al* 2012 & 2016). The methodological proposal I am making is along the same lines, but also considering group aspects.

There are three possible targets for what I called evolutionary social science at the beginning of the dissertation: the sociality of our ancestors, the social psychology we have inherited, and the forms of sociality as they manifest in contemporary societies. I have implicitly argued that there are reasons to approach the sociality of our ancestors holistically, in three dimensions relevant to evolution. (1) Some of the evolving social behavioural traits are likely to have been interactionist traits or traits of a collective. (2) The development of behavioural traits was under the influence of cultural factors, whether they were dependent on specific individuals (forming a developmental interaction group) or unified the entire collective regarding the given developmental factor. (3) Group selection (both interactive and collective) was a relevant level of selection, whether facilitated by the connections in the proximate or developmental dimension, or in the evolutionary dimension only, because of the group structure. Consequently, when we approach the social behaviour of our ancestors from an evolutionary functionalist perspective, we should not assume that the behavioural traits are individual adaptations. We should approach them through what function they have in group living, as a part of group "design".

As discussed in chapter 4, there are two strategies in applying evolutionary methodology to understanding contemporary human sociality: evolutionary psychology and evolutionary anthropology. Evolutionary psychology depends on assumptions about the behaviour that the psychology relates to, especially regarding social

behaviour. Behaviour gives psychology its evolutionary function. I made a clear distinction between psychological traits and behavioural traits in chapter 5: they are distinct even if coupled; behavioural traits and psychological traits are decoupled: any psychological trait is connected to multiple behavioural traits, and the other way around; and, finally, some behavioural traits should be understood as holistic traits. Even if evolutionary psychology is only interested in psychological makeup (a proximate-level individualist choice of perspective), and in particular the innate part of it (a developmental-level individualist choice of perspective), some of the relevant behavioural traits that determine the selective function (in the EEA) may be holistic, these traits may have had a holistic cultural basis (which may have changed), and group selection may have been important in the selection of the trait for various reasons. In other words, evolutionary psychology cannot be done independently of holistic evolutionary historical considerations which locate the evolution of the psychology in its selective context.

Evolutionary anthropology studies human behaviour directly and often studies it as adaptive to current conditions. As discussed before, this does not necessarily require an adaptation process to the current conditions, but the evolutionary functionalist methodology here may be ahistorically explanatory, as discussed in chapter 3. This assumes that the combination of evolved psychology, culture, and creative human thinking enables humans to adapt to the ecological conditions. In the case of anthropology, all the arguments given for holism apply directly.

10. Conclusion

The question I have now given a lengthy answer to was: given humans are social beings who evolved in groups, should we understand the evolutionary functionality of human social behaviour from the individual or group perspective? In other words: if we give an evolutionary explanation for social behaviour, or apply adaptationist heuristics in our attempts to understand human sociality, should we adopt individualistic or holistic approaches and frameworks?

I started the discussion by distinguishing three causal dimensions that are relevant to an evolutionary explanation: proximate, developmental, and evolutionary proper. I also distinguished two different senses in which the causal factors can be supra-individual or holistic: they may exist on the group level of biological organization in terms of the concrete organization of individuals in groups (collectivism), or they may be non-reducible properties of the interactions of individuals (interactionism). Interactions bind individuals into interaction groups with respect to the interaction, while collectives are the fixed groups of individuals living together.

My first substantial issue was to define what an evolutionary explanation is, from the contrastive-counterfactual and mechanistic point of view, including how the different dimensions are connected. I then argued that there are three defendable forms of *evolutionary functionalism* (or adaptationism) that make different claims and different assumptions. Any of them may be an adequate approach to human social behaviour in certain cases. *Current use functionalism* is an analysis of evolutionary functionality of behaviour in the current environment without making any assumption about its evolution. It is purely descriptive and has minimal explanatory power. I interpreted *behavioural ecology* mostly as a project of this kind. *Historical explanatory functionalism* is evolutionary functional analysis for historical purposes: it analyses traits as historical adaptations. *Ahistorical explanatory functionalism* uses adaptationist thinking as a guide to discover the proximate and developmental functionality of the organism. I argued

that this form of evolutionary functionalism is distinct from both of the other forms. Importantly, it makes assumptions about natural selection having been a significant factor in evolution but does not need to assume any particular trait that is analysed as if it is or has been an adaptation. Most evolutionary psychology and evolutionary anthropology can be understood either as historical or ahistorical explanatory functionalism. The distinction does not exist in the literature, neither within evolutionary human sciences nor in the philosophical literature. Evolutionary human sciences are usually understood as historical, but it would be more charitable, and justified, to consider them ahistorical. Ahistorical interpretation would be sufficient for the role that the evolutionary considerations are given as a heuristic to discover proximate and developmental mechanisms. Whichever methodological choice is made, it does not affect the inidividualism issue in the proximate or evolutionary dimensions, but it does affect what assumptions need to be made in the developmental dimension, which is an issue for evolutionary psychology especially.

My main argument about the proximate dimension was that some social behaviour should be understood as *interactive traits*: they emerge in the interaction between the individuals, and it is these traits that are selected for. The individual behavioural dispositions are secondary; individual traits are selected by virtue of their enabling the individuals to participate in these interactions. Furthermore, the object of selection is the psychological makeup on the individual. Behaviour for which the psychological characteristics are selected, are selected for the consequences of the behaviour that they participate in producing. If we approach behaviour from the evolutionary functionalist perspective the way I proposed (as a general way to analyse behaviour from the evolutionary point of view), the proper objects of evolutionary explanation are the *individual psychological traits* and the *interactive behavioural traits*.

As part of making this argument, I discussed the folk-psychological understanding of what human behaviour is, and I argued that this is not a proper way to understand human behaviour from the

evolutionary point of view, although, at the same time, folk psychology is a part of the phenomenon under study, and we should understand the role it plays in social behaviour – which is primarily not to attribute psychological states but to represent agents' dispositions. I made the distinction between *psychological, agentive,* and *behavioural* descriptions of traits and applied this to the concept of altruism, distinguishing between *psychological, agentive,* and *behavioural altruism*. I then used *reciprocal altruism* to illustrate and develop further the point about interactive behavioural traits and individual psychological traits. The social interactions characterized by reciprocity are the interactive behavioural traits that determine the fitness benefits and get selected, while the psychological makeup that is selected includes whatever capacities are needed for participating in these traits. Consequently, the individual traits that are selected are not parts of the interactive traits. Psychological traits are selected based on their overall fitness effects during the individual's lifetime. The forms of interaction that the individual participates in compete against each other, and their totality determines the fitness consequences of the psychological traits.

The major consequence of this argument is that there may be supra-individual adaptations even if they are not properties of a collective, if the replication of these traits goes vertically through individuals only (no social learning or any other influence from other individuals in the development is assumed here), and if the selection processes are individualistic. A minor consequence is that the evolution of reciprocal altruism does not require group selection: the competition between interactive traits is between which interactions make the individuals fittest, and the competition between the individual psychological traits is individualist selection for the makeup that makes the individual fittest given the social interactive environment. There is, however, a significant supra-individual element in this process that cannot be reduced to individualist presuppositions – it is, however, about the proximate-level presuppositions. We must understand the evolutionary function of reciprocal altruism from a holistic

perspective. This also highlights the importance of the connection between the proximate and evolutionary dimensions.

I presented two major arguments about the developmental dimension. First was about the role of culture. Building upon the *Extended Synthesis*, I argued that social interactions that transmit traits can be individualist, interactionist, or collectivist. Human cultures are characterized by horizontal (non-individualist) transmission, but this does not entail directly that culture makes cultural groups units of replication. The interactionist alternative is most plausible. However, my main point was about developmental individualism. I argued for a concept of innateness that makes sense in the context of the *Extended Synthesis*, and that this is the concept through which we should understand nativism in psychology, including evolutionary psychology. Furthermore, I suggested that nativism is a methodological choice for defining explanatory interests. This, however, has consequences for how to interpret the results of the research and its constraints. Again, these are not positions within the literature (neither in evolutionary psychological nor in philosophical), but my methodological recommendation. At the same time, I pointed out that this qualification of the research object highlights the limitations of nativism and calls for alternative approaches in evolutionary psychology.

The proximate and developmental dimensions are connected. I discussed the interactionist traits in the proximate dimension without holistic developmental components. However, if the reproduction of social traits is cultural, this enforces the existence of interactionist traits: the forms of interaction do not solely depend on the individual capacities in the interaction of which the interactive traits emerge, but they can be reproduced directly though social interactions. The connection works in the other direction, too: the existence of culturally transmitted traits that make the participating individuals fitter creates selection for individual psychological characteristics with the function of participating in these interactions.

The last dimension I discussed was the evolutionary, by which I mean the causal dimension of purely evolutionary factors – natural

selection being the only one of interest here. The individualism issue in this dimension is whether natural selection only works at the level of individuals (in the human context), or whether there is group selection too. Group selection has two senses, MLS-1 and MLS-2, which makes the distinction between interactionist and collectivist sense of groups and explanatory holism analogical to that in the other two dimensions. I discussed the relationship between MLS and kin selection, which has been the main recent theoretical controversy about the relationship between individuals and groups along this dimension, but I mostly concentrated on the relationship between the evolutionary and other dimensions in the group selection controversies. Partly building on Elisabeth Lloyd's work on the issue, I argued that keeping the different dimensions apart but analysing their connections using the ideas that I have developed over the course of the dissertation clarifies many of the remaining problems and mutual misunderstandings in the debate – especially what it means to have a group adaptation. Having a group selection process guiding the evolution, having a group-level adapting trait, and having a group-dependent replication process are three different dimensions. The group level may be needed in any one of them without a corresponding need in the others. Yet at the same time the supra-individual processes at any level may contribute causally to the emergence of supra-individual processes at the other levels. Furthermore, it is important to make the distinction between collectivist and interactionist versions of holism. They are different (in each dimension), but it is easy to equivocate between them in theoretical debates.

Generally speaking, I have concentrated on the interactionist forms of holism. Full-blown collectivism in both developmental and proximate dimensions are basically assumptions that the interactionist processes and mechanisms have taken over the whole group in a uniform fashion. Collectivism is an extreme version of the same holistic ideas that I have been discussing as interactionism. At the same time, there is a qualitative difference that emerges from this as the group becomes uniform. MLS-1 group selection can work on both the

interaction groups and collectives at the same time. MLS-2 group selection takes place between collective groups, but it does not require collectivism of the evolving trait in other dimensions. MLS-2 selection may enforce interactive traits as well if they make a difference between the collectives within which they exist. In short, interactionist processes in different dimensions reinforce each other, and interactionism may move towards collectivism in each dimension. Furthermore, if the evolutionary functions of different social traits are connected and cultural transmission connects the transmission of different traits (with whole traditions being transferred), this can make groups more and more like superorganisms.

To return to the main question: are human social behavioural traits adaptations of individuals or groups? The answer is likely: usually neither. If we take an evolutionary functionalist perspective on human social behaviour, both exclusively individualist and superorganism-evoking collectivist perspectives may be correct regarding some traits, but my methodological proposal is to approach it with a pluralistic toolkit of interactionist ideas along all three dimensions.

Bibliography

Aaltola, Elisa 2014. "Varieties of Empathy and Moral Agency", *Topoi* 33: 1–11.

Aaltola, Elisa 2018. *Varieties of Empathy: Moral Psychology and Animal Ethics*. Lanham: Rowman & Littlefield.

Achinstein, Peter 1983. *The Nature of Explanation*. Oxford: Oxford University Press.

Adriaens, Pieter R. & Andreas De Block (eds.) 2011. *Maladapting Minds: Philosophy, Psychiatry, and Evolutionary Theory*. Oxford: Oxford University Press.

Ågren, J. Arvid 2016. "Selfish genetic elements and the gene's-eye view of evolution", *Current Zoology* 62: 659–665.

Alcock, John 1979. *Animal Behavior: An Evolutionary Approach*. Sunderland: Sinauer.

Alcock, John 2003. *The Triumph of Sociobiology*. Oxford: Oxford University Press.

Alexander, Richard 1987. *The Biology of Moral Systems*. Aldine De Gruyter, New York.

Alroy, John & Alexander Levine 1994. "Driving both ways: Wilson & Sober's conflicting criteria for the identification of groups as vehicles of selection", *Behavioral and Brain Sciences* 17: 608–610.

Alvard, Michael S. 2003. "The adaptive nature of culture". *Evolutionary Anthropology: Issues, News, and Reviews* 12:136–149.

Amundson, Ron 1994. "Two Concepts of Constraint: Adaptationism and the Challenge from Developmental Biology", *Philosophy of Science* 61: 556–578.

Amundson, Ron 2001. "Adaptation and Development: On the Lack of Common Ground". In Steven H. Orzack & Elliot Sober (eds.): *Adaptationism and Optimality*. Cambridge: Cambridge University Press, 303–334.

Amundson, Ron & George V. Lauder 1993. "Function without Purpose: The Uses of Causal Role Function in Evolutionary Biology", *Biology and Philosophy* 9: 443–469.

Andersen, Hanne 2016. "Collaboration, interdisciplinarity, and the epistemology of contemporary science", *Studies in History and Philosophy of Science* 56: 1-10

Andersen, Holly K. 2011. "Mechanisms, Laws, and Regularities", *Philosophy of Science* 78: 325–331.

Anderson, John R. 1990. *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, Michael L. 2007. "The Massive Redeployment Hypothesis and the Functional Typography of the Brain", *Philosophical Psychology* 20: 143–174.

Andrews, Kristin 2012. *Do Apes Read Minds? Toward a New Folk Psychology*. Cambridge, Ma: MIT Press.

Andrews, Kristin 2015a. "Pluralistic folk psychology and varieties of self-knowledge: an exploration", *Philosophical Explorations* 18: 282–296.

Andrews, Kristin 2015b. "Folk psychological spiral: explanation, regulation, and language", *The Southern Journal of Philosophy* 53, Spindel Supplement: 50–67.
Anscombe, Elizabeth 1967. *Intention*. 2nd edition (2000 reprint). Cambridge, MA: Harvard University Press.
Apperly, Ian 2010. *Mindreaders: The Cognitive Basis of Theory of Mind*. New York, NY: Psychology Press.
Apperly, Ian A., Elisa Back, Dana Samson, Lisa France 2008. "The cost of thinking about false beliefs: Evidence from adults' performance on a non-inferential theory of mind task", *Cognition* 106: 1093–1108.
Apperly, Ian A., Daniel J. Carroll, Dana Samson, Glyn W. Humphreys, Adam Qureshi & Graham Moffitt 2010. "Why are there limits on theory of mind use? Evidence from adults' ability to follow instructions from an ignorant speaker", *The Quarterly Journal of Experimental Psychology* 63: 1201–1217.
Apperly, Ian, Kevin Riggs, Andrew Simpson, Dana Samson & Claudia Chiavarino 2006. "Is Belief Reasoning Automatic?" *Psychological Science* 17: 841–844.
Apperly, Ian & Elizabeth Robinson 2003. "When can Children Handle Referential Opacity? Evidence for Systematic Variation in 5- and 6-year-old Children's Reasoning About Beliefs and Belief Reports", *Journal of Experimental Child Psychology* 85: 297–311.
Ariew, Andre 1996. "Innateness and canalization", *Philosophy of Science* 63: S19–S27.
Ariew, Andre 1999. "Innateness is canalization: In defense of a developmental account of innateness." In Valerie Gray Hardcastle (ed.): *Where Biology Meets Psychology: Philosophical Essays*. Cambridge, Ma: MIT Press.
Ariew, Andre 2006. "Innateness." In Mohan Matthen & Christopher Stevens (eds.): *Handbook of the Philosophy of Biology*. Amsterdam: Elsevier.
Asquith, Peter & Philip Kitcher (eds.) 1985. *PSA 1984*. East Lansing: Philosophy of Science Association.
Atran, Scott 1990. *Cognitive Foundations of Natural History: Towards an Anthropology of Science*. Cambridge: Cambridge University Press.
Atran, Scott 2002. *In Gods We Trust: The Evolutionary Landscape of Religion*. Oxford: Oxford University Press.
Atran, Scott 2005, "Adaptationism for Human Cognition: Strong, Spurious or Weak?" *Mind & Language* 20: 39–67.
Atran, Scott & Douglas Medin 2008. *The Native Mind and the Cultural Construction of Nature*. Cambridge: MIT Press.
Atran, Scott, Douglas Medin, E. Lynch, V. Vaparansky, U.E. Edilberto & P. Sousa 2001. "Folkbiology doesn't come from folkpsychology: evidence from Yukatek Maya in cross-cultural perspective", *Journal of Cognition and Culture* 1: 3–42.

Atran, Scott, Douglas Medin & Norbert O. Ross 2005. "The Cultural Mind: Environmental Decision Making and Cultural Modeling Within and Across Populations", *Psychological Review* 112: 744–776.

Aunger, Robert (ed.) 2000. *Darwinizing Culture: The Status of Memetics as Science*. Oxford: Oxford University Press.

Avital, Eytan & Eva Jablonka 2000. *Animal Traditions: Behavioural Inheritance in Evolution*. Cambridge: Cambridge University Press.

Axelrod, Robert 1984. *The Evolution of Cooperation*. New York: Basic Books.

Back, Elisa & Ian Apperly 2010. "Two Sources of Evidence on the Non-automaticity of True and False Belief Ascription", *Cognition* 115: 54–70.

Badyaev, Alexander V. 2009. "Evolutionary significance of phenotypic accommodation in novel environments: an empirical test of the Baldwin effect", *Philosophical Transactions of the Royal Society B: Biological Sciences* 364: 1125–1141.

Badyaev, Alexander V. & Kevin P. Oh 2009. "Environmental induction and phenotypic retention of adaptive maternal effects", BMC *Evolutionary Biology* 8. 3 doi: 10.1186/1471-2148-8-3

Badlangana, N. Ludo, Justin W. Adams & Paul R. Manger 2009. "The giraffe (*Giraffa camelopardalis*) cervical vertebral column: a heuristic example in understanding evolutionary processes?" *Zoological Journal of the Linnean Society* 155: 736–757.

Baldwin, James Mark 1896. "A New Factor in Evolution", *The American Naturalist* 30: 441–451.

Bar-Anan, Yoav, Timothy D. Wilson & Ran R. Hassin 2010. "Inaccurate self-knowledge formation as a result of automatic behaviour", *Journal of Experimental Social Psychology* 46: 884–894.

Barclay, Pat 2013. "Strategies for cooperation in biological markets, especially for humans", *Evolution and Human Behavior* 34: 164–175.

Barclay, Pat 2016. "Biological markets and the effects of partner choice on cooperation and friendship", *Current Opinion in Psychology* 7: 33–38.

Bargh, John A. & Thanya L. Charthrand 1999. "The unbearable automaticity of being", *American Psychologist* 54: 462–479.

Barker, Gillian & Eric Desjardins & Trevor Pearce (eds.) 2014. *Entangled Life: Organism and Environment in Biological and Social Sciences*. Dordrecht: Springer.

Barkow, Jerome, Leda Cosmides & John Tooby (eds.) 1992. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York: Oxford University Press.

Barnard, Alan 2000. *History and Theory in Anthropology*. Cambridge: Cambridge University Press.

Baron-Cohen, Simon 1995. *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MA: Bradford/MIT Press.
Baron-Cohen, Simon, Helen Tager-Flusberg & Donald Cohen (eds.) 2000. *Understanding Other Minds* (2nd edition). Oxford: Oxford University Press.
Barrett, H. Clark, Leda Cosmides & John Tooby 2010. "Coevolution of cooperation, causal cognition and mindreading", *Communicative & Integrative Biology* 3: 522–524.
Barrett, Louise, Robin Dunbar & John Lycett 2002. *Human Evolutionary Psychology*. Basingstoke: Palgrave.
Barrett, Louise 2011. *Beyond the Brain: How body and environment shape animal and human minds*. Princeton University Press.
Barros, D. Benjamin 2008. "Natural Selection as a Mechanism", *Philosophy of Science* 75: 306–322.
Bar-Yosef, Ofer 2002. "The upper paleolithic revolution", *Annual Review of Anthropology* 31: 363–393.
Bateson, Patrick 1991. "Are there principles of behavioural development?" In P. Bateson (ed): *The Development and Integration of Behaviour: Essays in honour of Robert Hinde*. Cambridge: Cambridge University Press.
Bateson, Patrick & Matteo Mameli 2007. "The Innate and the Acquired: Useful Clusters or a Residual Distinction from Folk Biology?" *Developmental Psychobiology* 49: 818–831.
Batson, C. Daniel 1991. *The Altruism Question: Toward a Social-Psychological Answer*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
Batson, C. Daniel 1994. "Seeing the Light: What Does Biology Tell Us about Human Social Behavior?" *Behavioral and Brain Sciences* 17: 610–611.
Batson, C. Daniel 2000. "It's Been a Service… and a Disservice", *Journal of Consciousness Studies* 7: 207–256.
Batson, C. Daniel 2011. *Altruism in Humans*. Oxford University Press, Oxford.
Baumeister, Roy F. 2005. *The cultural animal: Human nature, meaning, and social life*. New York: Oxford University Press.
Beatty, John 1995. "The Evolutionary Contingency Thesis". In Wolters & Lennox (eds.): *Concepts, Theories, and Rationality in the Biological Sciences*. Pittsburgh: University of Pittsburgh Press, 45–81.
Bechtel, William 2006. *Discovering Cell Mechanisms: The creation of modern cell biology*. Cambridge University Press.
Bechtel, William 2008a. "Mechanisms in Cognitive Psychology: What Are the Operations?", *Philosophy of Science* 75: 983–994.

Bechtel, William 2008b. "Explanation: Mechanism, Modularity, and Situated Cognition". In Philip Robbins & Murat Aydede (eds.): *The Cambridge Handbook of Situated Cognition*. Cambridge: Cambridge University Press.

Bechtel, William 2011. "Mechanism and Biological Explanation", *Philosophy of Science* 78: 533–557.

Bechtel, William & Adele Abrahamsen 2005. "Explanation: a mechanistic alternative", *Studies in History and Philosophy of Biological and Biomedical Science* 36: 421–441.

Bechtel, William & Robert Richardson 1993. *Discovering Complexity*: *Decomposition and localization as strategies in scientific research*. Princeton, NJ: Princeton University Press.

Beck, Aaron T. 2008. "The evolution of the cognitive model of depression and its neurobiological correlates", *American Journal of Psychiatry* 165: 969–977.

Bedau, Mark 2010. "An Aristotelian Account of Minimal Chemical Life", *Astrobiology* 10: 1011–1020.

Bekoff, Marc 2002. "Animal Reflections", *Nature* 419: 255.

Bekoff, Marc & Jessica Pierce 2009. *Wild Justice: The Moral Lives of Animals*. Chicago: University of Chicago Press.

Bennett, Jonathan 1991. "Folk-psychological explanations". In John Greenwood (ed.): *The Future of Folk Psychology: Intentionality and Cognitive Science*. Cambridge: Cambridge University Press, 176–195.

Bentall, Richard P. 2003. *Madness Explained*. London: Penguin.

Bering, Jesse 2006. "The Folk Psychology of Souls", *Brain and Behavioral Sciences* 29: 453–498.

Berk, Laura 2013. *Child Development*. 9th Edition. New Jersey: Pearson.

Berlin, Brent 1992. *Ethnobiological Classification*. Princeton: Princeton University Press.

Bicchieri, Cristina 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

Birch, Jonathan 2009. "Irretrievably confused? Innateness in explanatory context", *Studies in History and Philosophy of Biological and Biomedical Sciences* 40: 296–301.

Birch, Jonathan 2012. "Collective action in the fraternal transitions", *Biology & Philosophy* 27: 363–380.

Birch, Jonathan 2014. "Gene Mobility and the Concept of Relatedness", *Biology and Philosophy* 29: 445–476.

Birch, Jonathan 2017. *The Philosophy of Social Evolution*. Oxford: Oxford University Press.

Birch, Jonathan 2019. "Are kin and group selection rivals or friends?" *Current Biology* 29: R433–R438.
Birch, Jonathan 2020. "Kin selection, group selection, and the varieties of population structure", *The British Journal for Philosophy of Science* 71: 259–286.
Birch, Jonathan & Samir Okasha 2015. "Kin selection and its critics", *BioScience* 65: 22–32.
Bjorklund, David F. & Anthony D. Pellegrini 2001. *The Origins of Human Nature: Evolutionary Developmental Psychology*, Washington, DC: American Psychological Association.
Blackburn, Simon 2001. *Ruling Passions*. Oxford: Clarendon Press.
Blackmore, Susan 1999. *The Meme Machine*. Oxford: Oxford University Press.
Boehm, Christopher 1999. *Hierarchy in the Forest: The Evolution of Egalitarian Behavior.* Cambridge: Harvard University Press.
Boehm, Christopher 2000a. "Conflict and the Evolution of Social Control", *Journal of Consciousness Studies* 7: 79–101.
Boehm, Christopher 2000b. "Group Selection in the Upper Palaeolithic", *Journal of Consciousness Studies* 7: 211–215.
Boehm, Christopher 2012. *Moral Origins: The Evolution of Virtue, Altruism and Shame*. New York: Basic Books.
Boesch, Christophe & Michael Tomasello 1998. "Chimpanzee and human cultures", *Current Anthropology* 39: 591–614.
Bogdan, Radu J. (ed.) 1991a. *Mind and Commonsense*. Cambridge, MA: Cambridge University Press.
Bogdan, Radu J. 1991b. "Common Sense Naturalized: The Practical Stance". In Radu J. Bogdan (ed.): *Mind and Commonsense*. Cambridge, MA: Cambridge University Press, 161–206.
Bogdan, Radu J. 1993. "The Architectural Nonechalance of Commonsense Psychology", *Mind and Language* 8: 189–205.
Bogdan, Radu J. 1997. *Interpreting Minds*. Cambrige, Ma: MIT Press.
Bogen, James 2005. "Regularities and Causality: Generalizations and Causal Explanations", *Studies in the History and Philosophy of Biology and Biomedical Sciences* 36: 397–420.
Bogen, James, and James Woodward 1988. "Saving the Phenomena", *Philosophical Review* 97: 303–352.
Boghossian, Paul 1989. "Content and Self-Knowledge", *Philosophical Topics* 17: 5–26.
Bolhuis, Johan 2005. "Function and mechanism in neuroecology: looking for clues", *Animal Biology* 55: 457–490.

Bolhuis, Johan, Gillian Brown, Robert Richardson & Kevin Laland 2011. "Darwin in Mind: New Opportunities for Evolutionary Psychology", *PLoS Biology* 9: e1001109. doi:10.1371/journal.pbio.1001109

Bolhuis, Johan & Luc-Alain Giraldeau 2009. "Introduction: Mechanisms of Animal Behaviour". In Johan Bolhuis & Simon Verhulst (eds.): *Tinbergen's Legacy: Function and Mechanism in Behavioral Biology*. Cambridge: Cambridge University Press.

Bolhuis, Johan & Simon Verhulst (eds.) 2009. *Tinbergen's Legacy: Function and Mechanism in Behavioral Biology*. Cambridge: Cambridge University Press.

Borrello, Mark 2010. *Evolutionary restraints: the contentious history of group selection*. Chicago: The University of Chicago Press.

Borgerhoff Mulder, Monique 1991. "Human behavioural ecology". In John Krebs & Nick Davies (eds.): *Behavioural Ecology: An Evolutionary Approach*. Oxford: Blackwell.

Borgerhoff Mulder, Monique & Ryan Schacht 2012. "Human behavioural ecology". In *Encycloepdia of Life Sciences*. Chichester: Wiley, 1–10.

Bouchard, Frédérick 2013. "How ecosystem evolution strengthens the case for functional pluralism". In Philippe Huneman (ed.): *Functions.* Springer. 83–95.

Bouchard, Frédérick & Philippe Huneman (eds.) 2013. *From Group to Individuals. Evolution and Emerging Individuality*. Cambridge, Ma: The MIT Press.

Bouchard, Frédérick & Alex Rosenberg 2004. "Fitness, Probability, and the Principles of Natural Selection", *British Journal for the Philosophy of Science* 55: 693–712.

Bourke, Andrew F.G. 2011. *Principles of Social Evolution.* Oxford University Press.

Bowlby, John 1969. *Attachments and Loss – Volume I: Attachement*. London: Hogarth Press.

Bowles, Samuel & Herbert Gintis 2011. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton & Oxford: Princeton University Press.

Boyd, Robert, Herbert Gintis, Samuel Bowles & Peter J. Richerson 2002. "The evolution of altruistic punishment". *PNAS* 100: 3531–3535.

Boyd, Robert & Peter J. Richerson 1985. *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Boyd, Robert & Peter J. Richerson 2000. "Memes: universal acid or a better mousetrap?" in Robert Aunger (ed.): *Darwinizing Culture*. Oxford: Oxford University Press, 143–162.

Boyd, Robert & Peter J. Richerson 2005. *The Origin and Evolution of Cultures*. Oxford: Oxford University Press.

Boyd, Robert Peter J. Richerson & Joseph Henrich 2011. "The cultural niche: Why social learning is essential for human adaptation", *Proceedings of the National Academy of Sciences USA* 108: 10918–10925.
Boyd, Robert & Joan B. Silk 2003. *How Humans Evolved*. 3rd edition. New York: W. W. Norton & Company.
Boyer, Pascal 2001. *Religion Explained: The Evolutionary Origins of Religious Thought*. New York: Basic Books.
Brace, C. Loring 1994. "The Consequences of Group Selection in a Domain without Genetic Input: Culture", *Behavioral and Brain Sciences* 17: 611–612.
Braddon-Mitchell, David, & Robert Nola (eds.) 2009. *Conceptual Analysis and Philosophical Naturalism*. Cambridge: MIT Press.
Bradie, Michael 1994. "Metaphors and mechanisms in vehicle-based selection theory", *Behavioral and Brain Sciences* 17: 612.
Braillard, Pierre-Alain & Christophe Malaterre (eds.) 2015. *Explanation in Biology: An Inquiry into the Diversity of Explanatory Patterns in the Life Sciences*. Dordrecht: Springer.
Brandon, Robert N. 1978 "Adaptation and Evolutionary Theory", *Studies in History and Philosophy of Science* 9: 181–206.
Brandon, Robert N. 1981 "Biological Teleology: Questions and Explanations", *Studies in History and Philosophy of Science* 12, 91–105.
Brandon, Robert N. 1982 "The levels of selection", *PSA 1982*, 315-323.
Brandon, Robert N. 1985. "Phenotype Plasticity, Cultural Transmission, and Human Sociobiology". In James H. Fetzer (ed.): *Sociobiology and Epistemology*. D. Reidel, 57–73.
Brandon, Robert N. 1988. "The levels of selection: a hierarchy of interactors", in H. Plotkin (ed.): *The Role of Behavior in Evolution*, 51–71.
Brandon, Robert N. 1990. *Adaptation and Environment*. Princeton: Princeton University Press.
Brandon, Robert N. 1996. *Concepts and Methods in Evolutionary Biology*. Cambridge University Press, Cambridge.
Brandon, Robert N. 2006. "The principle of drift: biology's First Law", *Journal of Philosophy* 102: 319–335.
Bratman, Michael 1987. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.
Bratman, Michael 1999. *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge: Cambridge University Press.

Brent, Lauren J.N., Julia Lehmann & Gabriel Ramos-Fernandez 2011. "Social Network Analysis in the Study of Nonhuman Primates: A Historical Perspective", *American Journal of Primatology* 73: 720–730.

Brigandt, Ingo 2005. "The Instinct Concept of the Early Konrad Lorenz," *Journal of the History of Biology* 38: 571–608.

Brookes, D., G. Horton, A. Van Heuvelen & E. Etkina 2005. "Concerning Scientific Discourse about Heat", in *2004 Physics Education Research Conference* (AIP Conference Proceedings) 790: 149–152.

Brown, Gillian, Thomas Dickins, Rebecca Sear & Kevin Laland 2011. "Evolutionary accounts of human behavioural diversity", *Philosophical Transactions of the Royal Society of London B* 366: 313–324.

Brown, Gillian & Peter Richerson 2013. "Applying evolutionary theory to human behaviour: past differences and current debates", *Journal of Bioeconomics* 16: 105–128.

Brüne, Martin & Ute Brüne-Cohrs 2006. "Theory of mind – evolution, ontogeny, brain mechanisms and psychopathology", *Neuroscience and Biobehavioral Reviews* 30: 437–455.

Bryant, Lauren, Anna Coffey, Daniel J. Povinelli & John R. Pruett, Jr 2013. "Theory of Mind experience sampling in typical adults", *Consciousness and cognition* 22: 697–707.

Buller, David 1998. "Etiological theories of function: A geographical survey", *Biology and Philosophy* 13: 505–527.

Buller, David 1999. "DeFreuding evolutionary psychology: adaptation and human motivation". In Valerie Gray Hardcastle (ed.): *Where Biology Meets Psychology: Philosophical Essays*, Cambridge, Ma: The MIT Press, 99–114.

Buller, David 2005. *Adapting Minds: Evolutionary Psychology and the Persistent Quest for Human Nature*. Cambridge: MIT Press.

Buller, David & Valerie Gray Hardcastle 2000. "Evolutionary Psychology, Meet Developmental Neurobiology: Against Promiscuous Modularity", *Brain and Mind* 1: 307–325.

Bunge, Mario 1979. *Treatise on Basic Philosophy, Vol. 4: Ontology II: A world of systems*. Dordrecht: Reidel.

Burge, Tyler 1979. "Individualism and the Mental", *Midwest Studies in Philosophy* 4: 73–121.

Burge, Tyler 2003. *Foundations of Mind: Philosophical Essays, Volume 2*. Oxford: Oxford University Press.

Burian, Richard 2001. "The dilemma of case studies resolved: the virtues of using case studies in the history and philosophy of science", *Perspectives on Science* 9: 383–404.
Burkhardt, Richard 1981. "On the emergence of ethology as a scientific discipline", *Conspectus of History* 1: 62–81.
Burkhardt, Richard 2005. *Patterns of Behavior: Konrad Lorenz, Niko Tinbergen, and the founding of Ethology*. Chicago: University of Chicago Press.
Buss, David 1995. "Evolutionary Psychology: A New Paradigm for Psychological Science", *Psychological Inquiry* 6: 1–30.
Buss, David (ed.) 2005. *The Handbook of Evolutionary Psychology*. Hoboken: John Wiley & Sons.
Buss, David 2014. *Evolutionary Psychology: The New Science of the Mind*. 4th edition. Boston: Allyn and Bacon.
Butler, Joseph 1729. *Fifteen Sermons Preached at the Rolls Chapel*. London: J. and J. Knapton.
Butterfield, Stephen A. & Ian A. Apperly 2013. "How to Construct a Minimal Theory of Mind", *Mind & Language* 28: 606–637.
Byrne, Richard W. 1997. "Machiavellian intelligence", *Evolutionary Anthropology* 5: 173–180.
Byrne, Richard W., Philip J. Barnard, Iain Davidson, Vincent M. Janik, William C. McGrew, Ádam Miklósi & Polly Wiessner 2004. "Understanding culture across species", *Trends in Cognitive Science* 8: 341–346.
Byrne, Richard W. & Andrew Whiten (eds.) 1988. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans*. Oxford: Oxford University Press.
Calcott, Brett 2009. "Lineage explanations: Explaining how biological mechanisms change", *British Journal for the Philosophy of Science* 60: 51–78.
Calcott, Brett 2013. "Why the proximate–ultimate distinction is misleading, and why it matters for understanding the evolution of cooperation." In Kim Sterelny, Richard Joyce, Brett Calcott & Ben Fraser (eds.): *Cooperation and Its Evolution*. Cambridge, Ma: The MIT Press, 249–264.
Calcott, Brett 2014. "Evolvability and Engineering", *Biology and Philosophy* 29: 293–313.
Calcott, Brett 2017. "Causal specificity and the instructive–permissive distinction", *Biology and Philosophy* 32: 481–505.
Calcott, Brett & Kim Sterelny (eds.) 2011. *The Major Transitions Revisited*. Cambridge, Ma: The MIT Press.

Call, Josep & Malinda Carpenter 2002. "Three sources of information in social learning." In Kerstin Dautenhahn & Chrystopher L. Nehaniv (eds.): *Imitation in animals and artifacts*. Cambridge, Ma: MIT Press, 211–228.

Call, Josep, Malinda Carpenter & Michael Tomasello 2005. "Copying results andcopying actions in the process of social learning: Chimpanzees and human children", *Animal Cognition* 8: 151–163.

Call, Josep, & Michael Tomasello 2008. "Does the chimpanzee have a theory of mind? 30 years later", *Trends in Cognitive Science* 12: 187–192.

Callender, Craig 2011. "Philosophy of Science and Metaphysics". In Steven French & Juha Saatsi (eds.): *The Continuum Companion to the Philosophy of Science*. New York: Continuum International Publishing Group.

Campbell, Donald T. 1969. "Ethnocentrism of disciplines and the fish-scale model of omniscience." In M. Sherif & C. W. Sherif (eds.), *Interdisciplinary Relationships in the Social Sciences*. Boston University Press. 328–348.

Campbell, Donald T. & John B. Gatewood 1994. "Ambivalently Held Group-optimizing Predispositions", *Behavioral and Brain Sciences* 17: 614.

Carey, Susan & Elizabeth Spelke 1996. "Science and core knowledge", *Philosophy of Science* 63: 515–533.

Carroll, Sean B. 2005. *Endless Forms Most Beautiful: The New Science of Evo Devo and the Making of the Animal Kingdom*. W. W. Norton & Company.

Carruthers, Peter 2002. "The cognitive functions of language", *Brain and Behavioral Sciences* 25: 657–726.

Carruthers, Peter 2006. *The Architecture of Mind*. Oxford: Oxford University Press.

Carruthers, Peter 2008. "An architecture for dual reasoning", in Jonathan St.B.T. Evans & Keith Frankish (eds.), *In Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press, 109–128.

Carruthers, Peter 2009. "How we know our own minds: the relationship between mindreading and metacognition", *Behavioral and Brain Sciences* 32: 121–138.

Carruthers, Peter 2011. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.

Carruthers, Peter 2013a. "The distinctively human mind: the many pillars of cumulative culture". In Hatfield & Pittman (eds.): *Evolution of Mind, Brain, and Culture*. University of Pennsylvania Press, 325–345.

Carruthers, Peter 2013b. "Mindreading in Infancy", *Mind and Language* 28: 141–172.

Carruthers, Peter & Peter K. Smith (eds.) 1996. *Theories of Theories of Mind*. Cambridge: Cambridge University Press.

Carruthers, Peter, S. Laurence, & Steven Stich (eds.) 2005. *The innate mind: Structure and contents*. New York: Oxford University Press.

Carruthers, Peter, S. Laurence & Stephen Stich (eds.) 2007. *The Innate Mind: Foundations and the future*. New York: Oxford University Press.
Cartwright, Nancy 1978. "Causal Laws and Effective Strategies", *Nous* 13: 419–437.
Cartwright, Nancy 1994. *Nature's Capacities and Their Measurements*. Oxford University Press.
Cartwright, Nancy 2004. "Causation: One word, many things", *Philosophy of Science* 71: 805–819.
Castro, Laureano, José M. Serrano, & Miguel A. Toro 1998. "Conceptual capacity to categorize and the evolution of altruism", *Journal of Theoretical Biology* 192: 561–565.
Cavalli-Sforza, Luigi Luca 2000. *Genes, Peoples and Languages*. New York: North Point Press.
Cavalli-Sforza, Luigi Luca & Marcus Feldman 1981. *Cultural Transmission and Evolution*. Princeton: Princeton University Press.
Cela-Concode, Camilo, & Francisco Ayala 2007. *Human Evolution: Its Trails from the Past*. Oxford: Oxford University Press.
Chagnon, Naopelon & William Irons 1979. *Evolutionary Biology and Human Social Behavior: An Anthropological Perspective*. North Scituate, MA: Duxbury Press.
Chakroff, Alek & Liane Young 2014. "The prosocial brain". In: Laura M. Padilla-Walker & Gustavo Carlo (eds.): *Prosocial Development: A Multidimensional Approach*. Oxford: Oxford University Press, 90–111.
Chapman, Royal N. 1928. "The quantitative analysis of environmental factors", *Ecology* 9: 111–112.
Chomsky, Noam 1959. "A Review of B. F. Skinner's Verbal Behavior", *Language* 35: 26–57.
Chomsky, Noam 1980. *Rules and Representations*. New York: Columbia University Press.
Chomsky, Noam 1986. *Knowledge of Language: Its Nature, Origin and Use*. New York: Praeger.
Chomsky, Noam 1993. *Language and Thought*. London: Moyer Bell.
Chomsky, Noam 2000. *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.
Chudek, Maciek, Wanying Zhao & Joseph Henrich 2013. "Culture-gene coevolution, large-scale cooperation, and the shaping of human social psychology." In Kim Sterelny, Richard Joyce, Brett Calcott & Ben Fraser (eds.): *Cooperation and Its Evolution*. Cambridge, Ma: The MIT Press, 425–458.
Churchland, Paul M. 1981. "Eliminative Materialism and the Propositional Attitudes", *Journal of Philosophy* 78: 67–90.

Churchland, Paul M. 1989. "Folk Psychology and the Explanation of Human Behavior". In James E. Tomberlin (series editor): *Philosophical Perspectives* Vol. 3: Philosophy of Mind and Action Theory. Ridgeview Publishing Company, 225–241.

Churchland, Patricia S. 1986. *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. Cambridge, Massachusetts: The MIT Press.

Churchland, Patricia S. 2011. *Braintrust: What Neuroscience Tells Us about Morality*. Princeton University Press.

Churchland, Patricia S. 2019. *Conscience: The Origins of Moral Intuition*. W. W. Norton & Company

Churchland, Paul M. & Patricia S. Churchland 1998. *On the Contrary: Critical Essays 1987-1997*. Cambridge, Massachusetts: The MIT Press.

Clatterbuck, Hayley 2015. "Chimpanzee Mindreading and the Value of Parsimonious Mental Models", *Mind & Language* 30: 414–436.

Clark, Andy & David Chalmers 1998. "The extended mind", *Analysis* 58: 7–19.

Clarke, Ellen 2014. "Origins of evolutionary transitions", *Journal of Biosciences* 39: 303–317.

Clarke, Ellen 2016. "A levels-of-selection approach to evolutionary individuality", *Biology & Philosophy* 31: 893–911.

Cocker, Mark & Richard Mabey 2005. *Birds Britannica*. London: Chatto & Windus.

Cohen, L. Jonathan 1992. *An Essay on Belief and Acceptance*. Oxford: Oxford University Press.

Corballis, Michael C., & Stephen E. G. Lea (eds.) 1999. *The Descent of Mind: Psychological Perspectives on Hominid Evolution*. Oxford: Oxford University Press.

Cosmides, Leda 1989. "The logic of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task", *Cognition* 21: 187–276.

Cosmides, Leda & John Tooby 1987. "From evolution to behaviour: evolutionary psychology as the missing link". In John Dupré (ed.): *The Latest of the Best: Essays on Evolution and Optimality*. Cambridge, MA: MIT Press.

Cosmides, Leda & John Tooby 1992. "Cognitive adaptations for social exchange". In: Jerome Barkow, Leda Cosmides & John Tooby (eds.): *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford: Oxford University Press.

Cosmides, Leda & John Tooby 1996. "Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty", *Cognition* 58, 1–73.

Cosmides, Leda & John Tooby (1997): "Evolutionary Psychology: A Primer". http://www.psych.ucsb.edu/research/cep/primer.html
Cosmides, Leda & John Tooby 2000. "Evolutionary Psychology and the Emotions". In M. Lewis & J. M. Haviland-Jones (eds.): *Handbook of Emotions*. New York: Guilford.
Cosmides, Leda & John Tooby 2005a. "Conceptual foundations of evolutionary psychology." In David Buss (ed.): *The Handbook of Evolutionary Psychology*. Hoboken: John Wiley & Sons, 5–67.
Cosmides, Leda & John Tooby 2005b. "Neurocognitive adaptations designed for social exchange". In Buss (ed.): *The Handbook of Evolutionary Psychology*. Hoboken: John Wiley & Sons, 584–627.
Cowie, Fiona 1999. *What's Within? Nativism Reconsidered*. Oxford: Oxford University Press.
Crane, Tim 1995. "The mental causation debate", *Proceedings of the Aristotelian Society* (Supplementary) 69: 211–236.
Craver, Carl F. 2001. "Role Functions, Mechanisms, and Hierarchy", *Philosophy of Science* 68: 53–74.
Craver, Carl F. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Clarendon Press.
Craver, Carl F. 2013. "Functions and Mechanisms: A Perspectivalist Account". In Philippe Huneman (ed.): *Functions*. Dordrecht: Springer.
Craver, Carl F. 2015. "Levels". In Thomas Metzinger & Jennnifer M. Windt (eds.). *Open MIND* 26: 1–26.
Crews, David & Ton Groothuis 2005. "Tinbergen's fourth question, ontogeny: sexual and individual differentiation", *Animal Biology* 55: 343–370.
Cronk, Lee 1994. "Group selection's new clothes", *Behavioral and Brain Sciences* 17: 615–617.
Cronk, Lee 1999. *That complex whole*. Boulder: Westview Press.
Cronk, Lee, Napoleon Chagnon & William Irons (eds.) 2000. *Adaptation and Human Behavior: An Anthropological Perspective*. Hawthorne, NY: Aldine de Gruyter.
Csibra, Gergely & Gergely György 2006. "Social learning and social cognition: the case for pedagogy", in Yuko Munakata & Mark H. Johnson (eds): *Processes of Change in Brain and Cognitive Development*. Proceedings of Attention and Performance XXI. Oxford: Oxford University Press, 249–274.
Csibra, Gergely & Gergely György 2011. "Natural pedagogy as evolutionary adaptation", *Philosophical Transactions of Royal Society, B: Biological Sciences* 366: 1149–1157.
Cummins, Robert 1975. "Functional Analysis", *Journal of Philosophy* 72: 741–765.

Cummins, Robert 1983. *The Nature of Psychological Explanation*. Cambridge, Ma: The MIT Press.
Cummins, Robert 1996. *Representations, Targets, and Attitudes*. Cambridge, Ma: The MIT Press.
Cushman, Fiery 2013. "The role of learning in punishment, prosociality, and human uniques." In Kim Sterelny, Richard Joyce, Brett Calcott & Ben Fraser (eds.): *Cooperation and Its Evolution*. Cambridge, Ma: The MIT Press, 333–372.
Cuthill, Innes 2005. "The study of function in behavioural ecology", *Animal Biology* 55: 399–417.
Daly, Martin & Margo Wilson 1999. "Human evolutionary psychology and animal behaviour", *Animal Behaviour* 57: 509–519.
Damuth, John A. & I. Lorraine Heisler 1988. "Alternative formulations of multilevel selection", *Biology and Philosophy* 3: 407–430.
Dancy, Jonathan 2000. *Practical Reality*. Oxford: Oxford University Press.
Danielson, Peter (ed.) 1998. *Modeling rationality, morality and evolution*. Oxford: Oxford University Press.
Darden, Lindley 2006. *Reasoning in Biological Discoveries*. Cambridge: Cambridge University Press.
Darwin, Charles 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.
Darwin, Charles 1871. *The Descent of Man and Selection in Relation to Sex*. John Murray, London.
Darwin, Charles 1872. *The Expression of the Emotions in Man and Animals*. London: John Murray.
Davidson, Donald 1963. "Actions, Reasons and Causes", *Journal of Philosophy* 60: 685–700.
Davidson, Donald 1970. "Mental Events". In Lawrence Foster and J. W. Swanson (eds.): *Experience and Theory*. Duckworth, London.
Davidson, Donald 1974. "Psychology as Philosophy". In S.C. Brown (ed.) *Philosophy of Psychology*. London: Palgrave Macmillan, 41–52.
Davidson, Donald 1978. "Intending". In Yirmiaku Yovel (ed.): *Philosophy of History and Action*. D. Reidel and the Magnes Press, 41–60.
Davies, Martin & Tony Stone (eds.) 1995. *Folk Psychology: The Theory of Mind Debate*. Cambridge: Blackwell.
Davies, Paul Sheldon, James H. Fetzer & Thomas R. Foster 1995. "Logical reasoning and domains specificity: a critique of the social exchange theory of reasoning", *Biology and Philosophy* 10: 1–37.
Dawkins, Richard 1976. *The Selfish Gene*. Oxford: Oxford University Press.

Dawkins, Richard 1982. *The Extended Phenotype*. Oxford: Oxford University Press.
Dawkins, Richard 1983. "Universal Darwinism". In D. S. Bendall (ed.): *Evolution from Molecules to Man*. Cambridge: Cambridge University Press, 403–425.
Dawkins, Richard 1984. "Replicator Selection and the Extended Phenotype". In Elliot Sober (ed.): *Conceptual Issues in Evolutionary Biology*. Cambridge, MA: The MIT Press, 125–141.
Dawkins, Richard 1986. *Blind Watchmaker*. Essex: Longman.
Dawkins, Richard 1989 [1976]. *The Selfish Gene*, second edition. Oxford: Oxford University Press.
Deacon, Terrence 1997. *The Symbolic Species: The Co-evolution of Language and the Brain*. W.W. Norton & Co.
Dediu, Dan 2015. *An Introduction to Gentics for Language Scientists: Current concepts, methods, and findings*. Cambridge: Cambridge Univeristy Press.
Dennett, Daniel C. 1971. "Intentional Systems", *The Journal of Philosophy* 68: 87–106.
Dennett, Daniel C. 1978. *Brainstorms*, Cambridge, MA: MIT Press.
Dennett, Daniel C. 1984. *Elbow Room. The Varieties of Free Will Worth Wanting*. Cambridge, Ma: MIT Press.
Dennett, Daniel C. 1987. *The Intentional Stance*. Cambridge, Ma: MIT Press.
Dennett, Daniel C. 1991a. *Consciousness Explained*. Boston: Little, Brown & Company.
Dennett, Daniel C. 1991b. "Two contrasts: folk craft versus folk science, and belief versus opinion". In John Greenwood (ed.): *The Future of Folk Psychology: Intentionality and Cognitive Science*. Cambridge: Cambridge University Press, 135–148.
Dennett, Daniel C. 1995. *Darwin's Dangerous Idea*. *Evolution and the Meanings of Life.* New York: Simon & Schuster.
Denniston, C. 1978. "An Incorrect Definition of Fitness Revisited", *Annals of Human Genetics* 42: 77–85.
DePaul, Michael & William Ramsey 1998. *Rethinking Intuition*. Lanham, MD: Rowman & Littlefield.
Depew, David J. & Bruce H. Weber 1996. *Darwinism Evolving: Systems Dynamics and the Genealogy of Natural Selection*. Cambridge, MA: The MIT Press.
DesAutels, Lane 2011. "Against Regular and Irregular Characterizations of Mechanisms", *Philosophy of Science* 78: 914–925.
DesAutels, Lane 2016. "Natural selection and mechanistic regularity", *Studies in History and Philosophy of Biological and Biomedical Sciences* 57: 13–23.
Devaine, Marie, Guillaume Hollard & Jean Daunizeau 2014. "Theory of Mind: Did Evolution Fool Us?" *PLOS ONE* 9: e87619. https://doi.org/10.1371/journal.pone.0087619

Devitt, Michael 2008. "Resurrecting Biological Essentialism", *Philosophy of Science* 75: 344–382.
Devitt, Michael & Kim Sterelny 1987. *Language and Reality*. Oxford: Blackwell.
de Waal, Frans 1982. *Chimpanzee Politics*. London: Cape.
de Waal, Frans 1989. *Peace-Making among Primates*. Harvard University Press, Cambridge MA.
de Waal, Frans 1996. *Good Natured*. Harvard University Press, Cambridge MA.
de Waal, Frans 2007. "The 'Russian doll' model of empathy and imitation", in S. Bråten (Ed.): *Advances in consciousness research vol. 68: On being moved: From mirror neurons to empathy*. Amsterdam, Netherlands: John Benjamins Publishing Company, 49–69.
de Waal, Frans 2009. *The Age of Empathy: Nature's Lessons for a Kinder Society*. Broadway Books.
de Waal, Frans 2012. *The Bonobo and the Atheist*. W. W. Norton & Company.
de Waal, Frans, & Kristin E. Bonnie 2009. "In tune with others: the social side of primate culture", in Kevin N. Laland & Bennett G. Galef eds): *The question of animal culture*. Cambridge: Harvard University Press, 19–39.
de Waal, Frans, Patricia Churchland & Telmo Pievani (eds.) 2014. *Evolved Morality: The Biology and Philosophy of Human Conscience*. Brill.
Diamond, Peter & Hannu Vartiainen 2012. *Behavioral Economics and Its Applications*. Princeton University Press.
Dixon, Geoffrey M. 1994. *Division Algebras: Octonions, Quaternions, Complex Numbers and the Algebraic Design of Physics*. Kluwer.
Dobzhansky, Theodosius 1973. "Nothing in biology makes sense except in the light of evolution", *American Biology Teacher* 35: 125–129.
Doherty, Martin 2009. *Theory of Mind. How Children Understand Others' Thoughts and Feelings*. Psychology Press.
Donald, Merlin 1998. "Mimesis and the executive suite: missing links in language evolution". In J. R. Hurford, M. Studdert-Kennedy, & C. Knight (Eds.): *Approaches to the Evolution of Language: Social and Cognitive Bases*. Cambridge, MA: Cambridge University Press, 44–67.
Donald, Merlin 2001. *A Mind So Rare: The Evolution of Human Consciousness*. New York: Norton.
Donald, Merlin 2018. "Key cognitive preconditions for the evolution of language", *Psychonomic Bulletin & Review* 24: 204–208.
Donohue, Kathleen 2004. "Density-dependent multilevel selection in the great lakes sea rocket", *Ecology* 85: 180–191.

Doris, John 2015. *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford: Oxford University Press.
Douglas, Heather 2009. *Science, Policy, and the Value Free Ideal*. Pittsburgh, PA: Pittsburgh University Press.
Dowe, Phil 2000. *Physical Causation*. Cambridge University Press.
Dray, William 1957. *Laws and Explanations in History*. Oxford: Oxford University Press.
Dray, William 1963. "The historical explanation of actions reconsidered". In S. Hook (ed.), *Philosophy and history: A symposium*. New York: New York University Press.
Dretske, Fred 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge: MIT Press.
Dugatkin, Lee Alan 1994. "Subtle Ways of Shifting the Balance in Favor of Between-Group Selection", *Behavioral and Brain Sciences* 17: 618–619.
Dunbar, Robin 1992. "Social Beahviour and Evolutionary Theory". In Steve Jones, Robert Martin & David Pilbeam (eds.): *The Cambridge Encyclopedia of Human Evolution*. Cambridge: Cambridge University Press, 145–147.
Dupré, John (ed.) 1987. *The Latest of the Best: Essays on Evolution and Optimality*. Cambridge, MA: MIT Press.
Dupré, John 1993. *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Harvard: Harvard University Press.
Dupré, John 2001. *Human Nature and the Limits of Science*. Oxford: Oxford University Press.
Dupré, John & M. O'Malley 2009. "Varieties of Living Things: Life at the Intersection of Lineage and Metabolism." *Philosophy and Theory in Biology* 1: 1-24.
Durham, William H. 1991. *Coevolution: Genes, Culture, and Human Diversity*. Palo Alto: Stanford University Press.
Durkheim, Émile 1895 [1982]. *Les Règles de la méthode sociologique*. Translated as *The Rules of Sociological Method*, W. D. Hall (trans.). Glencoe, IL: The Free Press.
Duval, Céline, Pascale Piolino, Alexandre Bejanin, Francis Eustache & Béatrice Desgranges 2011. "Age effects on different components of theory of mind", *Consciousness and Cognition* 20: 627–642.
Earnshaw, Eugene 2015. "Evolutionary forces and the Hardy–Weinberg equilibrium", *Biology and Philosophy* 30: 423–437.
Easton, Alexander & Nathan Emery (eds.) 2012. *The Cognitive Neuroscience of Social Behaviour*. Hove and New York: Psychology Press.
Eisenberg, Nancy & Tracy L. Spinard 2014. "Multidimensionality of Prosocial Behavior: Rethinking the Conceptualization and Development of Prosocial

Behavior". In: Laura M. Padilla-Walker & Gustavo Carlo (eds.): *Prosocial Development: A Multidimensional Approach*. Oxford: Oxford University Press, 17–40.

Elman, Jerry, Elizabeth Bates, Mark Johnson, Annette Karmiloff-Smith, Domenico Parisi & Kim Plunkett 1997. *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge: MIT Press.

Ellis, Brian 2001. *Scientific Essentialism*. Cambridge: Cambridge University Press.

Emery, Nathan 2012. "The evolution of social cognition". In Alexander Easton & Nathan Emery (eds.): *The Cognitive Neuroscience of Social Behaviour*. Hove and New York: Psychology Press, 115–156.

Emery, Nathan & Nicola Clayton 2004. "The Mentality of Crows: Convergent Evolution of Intelligence in Corvids and Apes", *Science* 306: 1903–1907.

Enc, Berent 1995. "Units of behavior", *Philosophy of Science* 62: 523–542.

Endler, John 1986. *Natural Selection in the Wild*. Princeton: Princeton University Press.

Enfield, N.J., & Stephen Levinson (eds.) 2006. *Roots of Human Sociality: Culture and cognition in interaction*. New York: Berg.

Epstein, Seymour 1994. "Integration of the cognitive and the psychodynamic unconscious", *American psychologist* 49: 709.

Ereshefsky, Marc 2007. "Psychological categories as homologies: lessons from ethology", *Biology and Philosophy* 22: 659–674.

Ereshefsky, Marc 2012. "Homology thinking", *Biology and Philosophy* 27: 381–400.

Eronen, Markus I. 2013. "No levels, no problems: Downward causation in neuroscience". *Philosophy of Science* 80: 1042–1052.

Eronen, Markus I. 2015. "Levels of organization: A deflationary account". *Biology and Philosophy* 30: 39–58.

Evans, Jonathan St. B.T. & David E. Over 1996. *Rationality and Reasoning*. Psychology Press.

Evans, Jonathan S.B.T., & Keith E. Stanovich 2013. "Dual-process theories of higher cognition: advancing the debate", *Perspectives on psychological science* 8: 223–241.

Fehr, Ernst & Urs Fischbacher 2004. "Social norms and human cooperation". *TRENDS in Cognitive Sciences* 8: doi:10.1016/j.tics.2004.02.007

Fehr, Ernst & Simon Gächter 2000. "Cooperation and Punishment in Public Goods Experiments", *The American Economic Review* 90: 980–994.

Fetzer, James H. (ed.) 1985. *Sociobiology and Epistemology*. Dordrecht: D. Reidel.

Fisher, Daniel C. 1985. "Evolutionary morphology: beyond the analogous, the anecdotal and the ad hoc", *Paleobiology* 11: 120–138.

Fisher, Helen 2016. *Anatomy of Love*. *A Natural History of Mating, Marriage, and Why We Stray*. W.W. Norton & Company.
Fisher, Ronald A. 1930. *The Genetical Theory of Natural Selection*. Oxford: Oxford University Press.
Fitch, W. Tecumseh, Marc D. Hauser & Noam Chomsky 2005. "The evolution of the language faculty: clarifications and implications", *Cognition* 97: 179-210.
Flack, Jessica C., & Frans B. M. de Waal 2000. "'Any animal whatever': Darwinian building blocks of morality on monkeys and apes", *Journal of Consciousness Studies* 7: 1–29.
Fodor, Jerry 1968. *Psychological explanation*, New York: Random House.
Fodor, Jerry 1974. "Special sciences (or: the disunity of science as a working hypothesis)", *Synthese* 28: 97–115.
Fodor, Jerry 1975. *The Language of Thought*. Cambridge: Harvard University Press.
Fodor, Jerry 1980. "Methodological Solipsism Considered as a Research Strategy in Cognitive Science", *Behavioral and Brain Sciences* 3: 63–73.
Fodor, Jerry 1981. *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, Ma: MIT Press.
Fodor, Jerry 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, Ma: MIT Press.
Fodor, Jerry 1987. *Psychosemantics*. Cambridge, MA: MIT Press.
Fodor, Jerry 1990. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.
Fodor, Jerry 2001. "Doing without *What's Within*: Fiona Cowie's critique of nativism," *Mind* 110: 99–148.
Fodor, Jerry, Thomas Bever and Merrill Garrett 1974. *The Psychology of Language*. New York: McGraw-Hill.
Foley, Robert 1992. "Studying Human Evolution by Analogy". In Steve Jones, Robert Martin & David Pilbeam (eds.): *The Cambridge Encyclopedia of Human Evolution*. Cambridge: Cambridge University Press, 335–340.
Foley, Robert, & Marta Mirazón Lahr 2003. "On stony ground: lithic technology, human evolution, and the emergence of culture", *Evolutionary Anthropology* 12: 109–122.
Forber, Patrick 2009. "Spandrels and a Pervasive Problem of Evidence", *Biololgy and Philosophy* 24: 247–266.
Forber, Patrick 2010. "Confirmation and explaining how possible", *Studies in History and Philosophy of Biological and Biomedical Sciences* 41: 32–40.
Forber, Patrick & Rory Smead 2015. "Evolution and the classification of social behavior", *Biology & Philosophy* 30: 405–421.

Formica, Vincent A., Corlett W. Wood, Phoebe Cook & Edmund Brodie III 2017. "Consistency of animal social networks after disturbance", *Behavioral Ecology* 28: 85–93.

Formica Vincent A., Joel W. Mcglothlin, Corlett W. Wood, Malcolm E. Augat, Rebecca E. Butterfield, Mollie E. Barnard & Edmund D. Brodie III 2011. "Phenotypic assortment mediates the effect of social selection in a wild beetle population", *Evolution* 65: 2771–2781.

Fox, Richard G. & Barbara J. King (eds.) 2002a. *Anthropology Beyond Culture*. Oxford: Berg.

Fox, Richard G. & Barbara J. King 2002b. "Introduction: Beyond culture worry", in Richard G. Fox & Barbara J. King (eds.): *Anthropology Beyond Culture*. Oxford: Berg. 1–19.

Frank, Robert 1988. *Passion within Reason: The Strategic Role of the Emotions*. New York: W.W. Norton.

Frank, Steven A. 1998. *Foundations of Social Evolution*. Princeton, NJ: Princeton University Press.

Frank, Steven A. 2013. "Natural selection. VII. History and interpretation of kin selection theory", *Journal of Evolutionary Biology* 26: 1151–1184.

Frankenhuis, Willem E., Karthik Panchanathan & H. Clark Barrett 2013. "Bridging developmental systems theory and evolutionary psychology using dynamic optimization". *Developmental Science* 16: 584–598.

Frankish, Keith 2004: *Mind and Supermind*. Cambridge: Cambridge University Press.

Frankish, Keith & Jonathan Evans (eds.) 2009: *In Two Minds*. Oxford: Oxford University Press.

Friedman, Michael 1974. "Explanation and Scientific Understanding", *Journal of Philosophy* 71: 5–19.

Futuyma, Douglas 1998. *Evolutionary Biology*, 3rd edition, Sunderland, MA: Sinauer Associates.

Futuyma, Douglas 2010. "Evolutionary constraint and ecological consequences", *Evolution* 64: 1965–1884.

Gajdon, Gyula K., Natasha Fijn & Ludwig Huber 2004. "Testing social learning in a wild mountain parrot, the kea (Nestor notabilis)", *Animal Learning & Behavior* 32: 62–71.

Galef, Bennett G. 1992. "The question of animal culture", *Human Nature* 3: 157–178.

Gallagher, Shaun 2007. "Logical and phenomenological arguments against simulation theory", in Daniel D. Hutto & Matthew Ratcliffe (eds.) 2007: *Folk Psychology Reassessed*. Springer, 63–78.

Gallagher, Shaun 2012. "Empathy, simulation and narrative", *Science in Context* 25: 355–381.
Gallotti, Mattia & Chris Frith 2013. "Social cognition in the we-mode", *Trends in Cognitive Science* 17: 160–165.
Gangestad, Steven, & Jeffry Simpson (eds.) 2007. *The Evolution of Mind: Fundamental Questions and Controversies*. New York: Guilford Press.
Gánti, Tibor 2003. *The Principles of Life*. Oxford: Oxford University Press.
Gärdenfors, Peter 1980. "A Pragmatic Approach to Explanations", *Philosophy of Science* 47: 404–423.
Gardner, Andy 2013a. "Ultimate explanations concern the adaptive rationale for organism design", *Biology and Philosophy* 28: 787–791.
Gardner, Andy 2013b. "Adaptation of individuals and groups." In Frédérick Bouchard & Philippe Huneman (eds.): *From Groups to Individuals*. Cambridge, Ma: The MIT Press, 99–116.
Gardner, Andy 2015a. "The genetical theory of multilevel selection", *Journal of Evolutionary Biology* 28: 305–319.
Gardner, Andy 2015b. "Group selection versus group adaptation", *Nature* 524(7566): E3-E4. Doi: http://dx.doi.org/10.1038/nature14596.
Gardner, Andy, Stuart A. West, Geoff Wild 2011. "The genetical theory of kin selection", *Journal of Evolutionary Biology* 24: 1020–1043. http://dx.doi.org/10.1111/j.1420-9101.2011.02236.x.
Garson, Justin 2016. *A Critical Overview of Biological Functions*. Springer.
Gauvain, Mary & Robert L. Munroe 2012. "Cultural Change, Human Activity, and Cognitive Development". *Human Development* 55: 205–228.
Gawronski, Bertram, Wilhelm Hofmann & Christopher J. Wilbur 2006: "Are 'implicit' attitudes unconscious?" *Consciousness and Cognition* 15 485–499.
Geertz, Clifford 1973. *The Interpretation of Cultures: Selected Essays*. New York: Basic Books.
Gelman, Susan 2003. *The Essential Child: Origins of Essentialism in Everyday Thought*. New York: Oxford University Press.
Gendler, Tamar Szabó 2008. "Alief and Belief", *Journal of Philosophy* 105: 634–663.
Gibson, Kathleen R. & Maggie Tallerman (eds.) 2011. *The Oxford Handbook of Language Evolution*. Oxford: Oxford University Press.
Giddens, Anthony 1986. *The Constitution of Society: Outline of the Theory of Structuration*. University of California Press.
Gigerenzer, Gerd 2007. *Gut feelings: The intelligence of the unconscious*. New York: Viking Penguin.

Gilbert, Margaret 1994. "Me, You, and Us: Distinguishing "Egoism", "Altruism," and "Groupism", *Behavioral and Brain Sciences* 17: 621–622.
Gilbert, Margaret 1996. *Living Together: Rationality, Sociality, and Obligation*. New York: Rowman and Littlefield.
Gilbert, Scott F., John Opitz & Rudolf Raff 1996. "Resynthesizing evolutionary and developmental biology", *Developmental Biology* 173: 357–72.
Gilbert, Scott F. 2001. "Ecological developmental biology: developmental biology meets the real world". *Developmental Biology* 233: 1–22.
Gilbert, Scott F. 2003. "Evo-Devo, Devo-Evo, and Devgen-Popgen", *Biology and Philosophy* 18: 347–352.
Gilbert, Scott F. 2007. *Developmental biology*. 8th edition. Sunderland, MA: Sinauer.
Gil-White, Fransesco 2001. "Are Ethnic Groups Biological Species to the Human Brain?" *Current Anthropology* 42: 515–54.
Gintis, Herbert 2000a. "Strong reciprocity and human sociality", *Journal of Theoretical Biology* 206: 169–179.
Gintis, Herbert 2000b. "Group selection and human prosociality", *Journal of Consciousness Studies* 7: 215–219.
Gintis, Herbert 2006. "Behavioral ethics meets natural justice". *Politics, Philosophy & Economics* 5: 5–32.
Gintis, Herbert 2007. "A framework for the unification of the behavioral sciences", *Behavioral and Brain Sciences* 30: 1–61.
Gintis, Herbert 2009. *Game Theory Evolving*, Second Edition. Princeton: Princeton University Press.
Glennan, Stuart S. 1996. "Mechanisms and the nature of causation", *Erkenntnis* 44: 49–71.
Glennan, Stuart S. 2002a. "Contextual Unanimity and the Units of Selection Problem", Philosophy of Science 69: 118–137.
Glennan, Stuart S. 2002b. "Rethinking Mechanistic Explanation", *Philosophy of Science* 69: 342–353.
Glennan, Stuart S. 2010a. "Mechanisms, Causes, and the Layered Model of the World", *Philosophy and Phenomenological Research* 81: 362–381.
Glennan, Stuart S. 2010b. "Mechanisms". In Helen Beebee, Christopher Hitchcock & Peter Menzies (eds.): *Oxford Handbook of Causation*, 315–325.
Glennan, Stuart S. 2015. "When is it mental?" *HumanaMente* 29: 141-161.
Godfrey-Smith, Peter 1993. "Functions: consensus without unity", *Pacific Philosophical Quarterly* 74: 196–208.
Godfrey-Smith, Peter 1996. *Complexity and the Function of Mind in Nature*.

Godfrey-Smith, Peter 2000. "On the theoretical role of 'genetic coding'", *Philosophy of Science* 67: 26–44. Cambridge: Cambridge University Press.
Godfrey-Smith, Peter 2001. "Three Kinds of Adaptationism". In Steven H. Orzack & Elliot Sober (eds.): *Adaptationism and Optimality*. Cambridge: Cambridge University Press, 335–357.
Godfrey-Smith, Peter 2005. "Folk Psychology as a Model", *Philosopher's Imprint* 5: 1–15.
Godfrey-Smith, Peter 2006a. "Mental Representation, Naturalism and Teleosemantics", in Papineau and Macdonald (eds.): *New Essays on Teleosemantics*. Oxford University Press, 42–68.
Godfrey-Smith, Peter 2006b. "Local interaction, multilevel selection, and evolutionary transitions", *Biological Theory* 1: 372–380.
Godfrey-Smith, Peter 2008a. "Reduction in real life". In Jakob Hohwy & Jesper Kallestrup (eds.): *Being Reduced: New Essays on Reduction, Explanation, and Causation*. Oxford: Oxford University Press.
Godfrey-Smith, Peter 2008b. "Varieties of population structure and the levels of selection", *British Journal for the Philosophy of Science* 59: 25–50.
Godfrey-Smith, Peter 2009. *Darwinian Populations*. Oxford: Oxford University Press.
Godfrey-Smith, Peter 2010. "Causal Pluralism", in Helen Beebee, Christopher Hitchcock & Peter Menzies (eds.), *Oxford Handbook of Causation*. Oxford: Oxford University Press, 326–337.
Godfrey-Smith, Peter 2013. "Darwinian Individuals." In Frédérick Bouchard and Philippe Huneman (eds.): *From Groups to Individuals: Perspectives on Biological Associations and Emerging Individuality*. Cambridge MA: MIT Press, 17–36.
Godfrey-Smith, Peter 2014. "Individuality and Life Cycles". In Thomas Pradeu & Alexandre Guay (eds.): *Individuals Across the Sciences*. Oxford: Oxford University Press.
Goldie, Peter 2007. "There are reasons and reasons". In Daniel D. Hutto & Matthew Ratcliffe (eds.) (2007): *Folk Psychology Reassessed*. Springer, 103–114.
Goldman, Alvin 1989. "Interpretation Psychologized", *Mind and Language* 4: 161–185.
Goldman, Alvin 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.
Goodman, Morris 1992. "Reconstructing Human Evolution from Proteins". In Steve Jones, Robert Martin & David Pilbeam (eds.): *The Cambridge Encyclopedia of Human Evolution*. Cambridge University Press, Cambridge, 307–312.

Goodnight, Charles J. 2012. "On multilevel selection and kin selection: Contextual analysis meets direct fitness", *Evolution* 67: 1539–1548.
Goodnight, Charles J. 2013a. "Defining the individual". In Frédéric Bouchard & Philippe Huneman (eds.): *From Groups to Individuals*. Cambridge, Ma: The MIT Press.
Goodnight, Charles J. 2013b. "On multilevel selection and kin selection: Contextual analysis meets direct fitness", *Evolution* 67: 1539–1548.
Goodnight, Charles J. 2015. "Multilevel selection theory and evidence: a critique of Gardner", *Journal of Evolutionary Biology* 28: 1734–1746.
Goodnight, Charles J. & Lori Stevens 1997. "Experimental studies of group selection: what do they tell us about group selection in nature?" *American Naturalist* 150: S59–S79.
Goodnight, Charles J., James M. Schwartz & Lori Stevens 1992. "Contextual analysis of models of group selection, soft selection, hard selection, and the evolution of altruism", *The American Naturalist* 140: 743–761.
Gopnik, Alison, & Andrew Meltzoff 1997. *Words, Thoughts and Theories*. Cambridge, Ma: The MIT Press.
Gordon, Robert 1986. "Folk Psychology as Simulation", *Mind and Language* 1: 158–171.
Gottlieb, Gilbert 1992. *Individual Development and Evolution*. Oxford: Oxford University Press.
Gottlieb, Gilbert 1997. *Synthesizing Nature–Nurture: Prenatal Roots of Instinctive Behavior*. Hillsdale, NJ: Lawrence Erlbaum Associactes.
Gould, Stephen Jay 1984. "Caring Groups and Selfish Genes". In Elliot Sober (ed.): *Conceptual Issues in Evolutionary Biology*. Cambridge, MA: The MIT Press, 119–124.
Gould, Stephen Jay 2002. *The Structure of Evolutionary Theory*, Cambridge, MA: Harvard University Press.
Gould, Stephen Jay & Richard C. Lewontin 1979. "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationst Programme", *Proceedings of Royal Society of London* 205: 581–598.
Gould, Stephen Jay & Elisabeth S. Vrba 1982. "Exaptation – A Missing Term in the Science of Form", *Paleobiology* 8: 4–15.
Grafen, Alan 1984. "Natural selection, kin selection and group selection", in John Krebs & Nick Davies (eds.): *Behavioural Ecology: An Evolutionary Approach*. Blackwell, 62–84.
Graham, Jesse, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik & Peter H. Ditto 2013. "Moral Foundations Theory: The pragmatic validity of moral pluralism", *Advances in Experimental Social Psychology* 47: 55–130.

Green, Sara 2014. "A Philosophical Evaluation of Adaptationism as a Heuristic Strategy", *Acta Biotheoretica* 62: 479–498.
Greenspan, Ralph 2008. "The origins of behavioral genetics", *Current Biology* 18: 192–198.
Greenwood, John (ed.) 1991. *The Future of Folk Psychology: Intentionality and Cognitive Science*. Cambridge: Cambridge University Press.
Grene, Marjorie & David Depew 2004. *The Philosophy of Biology: An Episodic History*. Cambridge: Cambridge University Press.
Gribbin, John & Jeremy Cherfas 2001. *The First Chimpanzee: In Search of Human Origins*. Penguin Books, London.
Griffiths, Paul E. 1993. "Functional Analysis and Proper Functions", *British Journal for Philosophy of Science* 44: 409–421.
Griffiths, Paul E. 1997. *What Emotions Really Are? The Problem of Psychological Categories*. Chicago: University of Chicago Press.
Griffiths, Paul E. 2001. "Genetic information: a metaphor in search of a theory". *Philosophy of Science* 68: 394– 412.
Griffiths, Paul E. 2002. "What is Innateness?" *The Monist* 85, 70–85.
Griffiths, Paul E. 2004. "Instinct in the '50s: The British Reception of Konrad Lorenz's Theory of Instinctive Behaviour," *Biology and Philosophy* 19, 609–631.
Griffiths, Paul E. 2006. "Function, Homology and Character Individuation", *Philosophy of Science* 73: 1–25.
Griffiths, Paul E. 2009. "The distinction between innate and acquired characters." In Edward N. Zalta (ed.): *Stanford Encyclopedia of Philosophy*. URL: http://plato.stanford.edu/archives/fall2009/entries/innate-acquired/
Griffiths, Paul E. & Russell D. Gray 1994. "Developmental Systems and Evolutionary Explanation", *Journal of Philosophy* 91: 277–304.
Griffiths, Paul E. & Russell D. Gray 1997. "Replicator II: Judgement Day". *Biology and Philosophy* 12: 471–92.
Griffiths, Paul E. & Russell D. Gray 2005. "Three Ways to Misunderstand Developmental Systems Theory". *Biology and Philosophy* 20: 417–25.
Griffiths, Paul E. & Edouard Machery 2008. "Innateness, canalisation and 'biologicizing the mind'", *Philosophical Psychology* 21: 397–414.
Griffiths, Paul E., Edouard Machery & Stefan Linquist 2009. "The vernacular concept of innateness", *Mind and Language* 24: 605–630.
Griffiths, Paul E. & Karola Stotz 2008. "Gene", in David L. Hull & Micahel Ruse (eds.), *The Cambridge Companion to the Philosophy of Biology*, Cambridge: Cambridge University Press, 85–102.

Griffiths, Paul E. & Karola Stotz 2013. *Genetics and Philosophy: An Introduction*. Cambridge: Cambridge University Press.
Griffiths, Paul E. & James Tabery 2013. "Developmental Systems Theory: What Does It Explain, and How Does It Explain It?" *Advances in Child Development and Behavior* 44: 65–94.
Hacking, Ian 1999. *Social Construction of What?* Cambridge, MA: Harvard University Press.
Haidt, Jonathan 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment", *Psychological Review* 108: 814–834.
Haidt, Jonathan 2012. *The Righteous Mind: Why Good People are Divided By Politics and Religion*. New York: Pantheon Books.
Haidt, Jonathan & F. Bjorklund 2008. "Social intuitionists answer six questions about moral psychology". In W. Sinnott Armstrong (ed.): *Moral Psychology. Vol.2: The Cognitive Science of Morality*. Cambridge, Ma: MIT Press, 181–217.
Haidt, Jonathan & Joseph Craig 2004. "Intuitive ethics: how innately prepared intuitions generate culturally variable virtues", *Daedalus* 133: 55–66.
Hakli, Raul, Kaarlo Miller & Raimo Tuomela 2010. "Two Kinds of We-Reasoning", *Economics & Philosophy* 26: 291–320.
Haldane, John 1932. *The Causes of Evolution*. London: Longmans, Green & Co.
Haldane, John 1955. "Population genetics", *New Biology* 18: 34–51.
Hall, Brian 2003. "Evo-Devo: evolutionary developmental mechanisms", *International Journal of Developmental Biology* 47: 491–495.
Hall, Ned 2004. "Two concepts of causation". In John Collins, Ned Hall & Laurie Paul (eds.), *Causation and Counterfactuals*. Cambridge, Ma: The MIT Press, 225–276.
Hamilton, Andrew & Jennifer Fewell 2013. "Groups, individuals, and the emergence of sociality: the case of division of labor". In Frédérick Bouchard & Philippe Huneman (eds.): *From Group to Individuals. Evolution and Emerging Individuality*. Cambridge, Ma: The MIT Press. 175–194.
Hamilton, Andrew, Nathan R. Smith & Matthew H. Haber 2009. "Social insects and the individuality. Thesis: cohesion and the colony as a selectable individual." In: Jürgen Gadau & Jennifer Fewell (eds.), *Organization of insect societies: from genome to sociocomplexity*. Cambridge: Harvard University Press, 570–596.
Hamilton, William D. 1964a. "The genetical evolution of social behaviour. I", *Journal of Theoretical Biology* 7: 1–16.
Hamilton, William D. 1964b. "The genetical evolution of social behaviour. II", *Journal of Theoretical Biology* 7: 17–52.

Hamilton, William D. 1970. "Selfish and spiteful behaviour in an evolutionary model", *Nature* 228: 1218–20.
Hardcastle, Valerie Gray (ed.) 1999. *Biology Meets Psychology: Constraints, Conjectures, Connections*. Cambridge, Ma: The MIT Press.
Hare, Brian & Shinya Yamamoto (eds.) 2015. *Bonobo Cognition and Behaviour*. Brill.
Hargreaves Heap, Shaun P. & Yanis Varoufakis 1995. *Game Theory: A Critical Introduction*. London and New York: Routledge.
Harman, Gilbert 2000. "Can Evolutionary Theory Provide Evidence against Psychological Hedonism?" *Journal of Consciousness Studies* 7: 219–221.
Harms, William & Brian Skyrms 2008. "Evolution of moral norms", in Michael Ruse (ed.): *The Oxford Handbook of Philosophy of Biology*. Oxford: Oxford University Press.
Harre, Rom, & Edward Madden 1975. *Causal Powers: Theory of Natural Necessity*. Blackwell.
Harvey, Paul H. & Mark D. Pagel 1991. *The Comparative Method in Evolutionary Biology*. Oxford: Oxford University Press.
Hatfield, Gary & Holly Pittman (eds.) 2013. *Evolution of Mind, Brain, and Culture*. University of Pennsylvania Press.
Hauser, Marc 2006. *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. New York: HarperCollins.
Hauser, Marc D., Charles Yang, Robert C. Berwick, Ian Tattersall, Michael J. Ryan, Jeffrey Watumull, Noam Chomsky & Richard C. Lewontin 2014. "The mystery of language evolution", *Frontiers in Psychology* 07 May 2014. doi.org/10.3389/fpsyg.2014.00401
Hawley, Patricia 2014. "Evolution, Prosocial Behavior, and Altruism: A Roadmap to Understanding Where the Proximate Models Meet Ultimate". In: Laura M. Padilla-Walker & Gustavo Carlo (eds.): *Prosocial Development: A Multidimensional Approach*. Oxford: Oxford University Press, 43–69.
Heberlein, Andrea S. & Ralph Adolphs 2012. "Functional anatomy of human social cognition". In Alexander Easton & Nathan Emery (eds.): *The Cognitive Neuroscience of Social Behaviour*. Hove and New York: Psychology Press, 157–194.
Heil, John, & Alfred Mele (eds.) 1993. *Mental Causation*, Oxford: Clarendon Press.
Hempel, Carl Gustav 1965. *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*. Free Press, New York.
Henderson, David 1993. *Interpretation and Explanation in the Human Sciences*. Albany: State University of New York Press.
Henderson, David 2005. "Norms, Invariance, and Explanatory Relevance", *Philosophy of the Social Sciences* 35: 1–15.

Henrich, Joseph 2001. "Cultural Transmission and the Diffusion of Innovations: Adoption Dynamics Indicate That Biased Cultural Transmission Is the Predominant Force in Behavioral Change", *American Anthropologist* 103: 992–1013.
Henrich, Joseph 2004a. "Cultural Group Selection, Coevolutionary Processes and Large-Scale Cooperation", *Journal of Economic Behavior & Organization* 53: 3–35.
Henrich, Joseph 2004b. "Demography and Cultural Evolution: How Adaptive Cultural Processes Can Produce Maladaptive Losses – The Tasmanian Case", *American Antiquity* 69: 197–214.
Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr & Herbert Gintis 2004. *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies.* Oxford: Oxford University Press.
Henrich, Joseph, Steven Heine & Ara Norenzayan 2010. "The weirdest people in the world?" *Behavioral and Brain Sciences* 33: 61–83.
Henrich, Joseph & Richard McElreath 2003. "The Evolution of Cultural Evolution", *Evolutionary Anthropology* 12: 123–135.
Henrich, Natalie, & Joseph Henrich 2007. *Why Humans Cooperate: A Cultural and Evolutionary Explanation*. Oxford: Oxford University Press.
Henry, Julie D., Louise H. Phillips, Ted Ruffman, Phoebe E. Bailey 2013. "A meta-analytic review of age differences in theory of mind", *Psychology and Aging* 28: 826–839.
Henshilwood, Christopher S. & Curtis W. Marean 2003. "The Origin of Modern Human Behavior: Critique of the Models and Their Test Implications", *Current Anthropology* 44: 627–651.
Heisler, I. Lorraine & John A. Damuth 1987. "A method for analyzing selection in hierarchically structured populations", *The American Naturalist* 130: 582–602.
Heyes Cecilia M. 1993. "Imitation, culture and cognition", *Animal Behaviour* 46: 999–1010.
Heyes Cecilia M. 1994. "Social learning in animals: categories and mechanisms", *Biological Review* 69: 207–231.
Heyes, Cecilia M. 2013. "What can imitation do for cooperation?" In Kim Sterelny, Richard Joyce, Brett Calcott & Ben Fraser (eds.): *Cooperation and Its Evolution*. Cambridge, Ma: The MIT Press. 312–331.
Hinde, Richard 1968. "Dichotomies in the study of development", in J. M. Thoday & A. S. Parkes (eds): *Genetic and Environmental Influences on Behaviour*. New York: Plenum.
Hindriks, Frank 2009. "Constitutive rules, language, and ontology". *Erkenntnis* 71: 253–275.

Hintikka, Jaakko 1988. "What is the Logic of Experimental Inquiry". *Synthese* 74: 173–190.
Hirschfeld, Lawrence A. 1995. "Do children have a theory of race?" *Cognition* 54, 209–252
Hirschfeld, Lawrence A. & Susan A. Gelman (eds.) 1994. *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge: Cambridge University Press.
Hitchcock, Christopher 2001. "A tale of two effects", *The Philosophical Review* 110: 361–396.
Hitchcock, Christopher 2003. "Of Humean Bondage", *British Journal for Philosophy of Science* 54: 1–25.
Hitchcock, Christopher & Joel Velasco 2014. "Evolutionary and Newtonian forces", *Ergo* 1. http://dx.doi.org/10.3998/ergo.12405314.0001.002
Hogan, Jerry 2005. "Causation: the study of behavioural mechanisms", *Animal Biology* 55: 323–341.
Hogan, Jerry & Johan Bolhuis 2005. "The development of behaviour: trends since Tinbergen (1963)", *Animal Biology* 55: 371–398.
Hogan, Jerry & Johan Bolhuis 2009. "Tinbergen's Four Questions and Contemporary Behavioral Biology". In Johan Bolhuis & Simon Verhulst (eds.): *Tinbergen's Legacy: Function and Mechanism in Behavioral Biology*. Cambridge: Cambridge University Press, 25–34.
Hogarth, Robin & Melvin W. Reder 1987. *Rational Choice: The Contrast between Economics and Psychology*. Chicago: University of Chicago Press.
Holsinger, Kent E. 1994. "Groups as vehicles and replicators: the problem of group-level adaptation", *Behavioral and Brain Sciences* 17: 626–623.
Hoppitt, William & Kevin N. Laland 2013. *Social Learning: An Introduction to Mechanisms, Methods, and Models*. Princeton University Press.
Horgan, Terrence & James Woodward 1985. "Folk Psychology is Here to Stay", *Philosophical Review* 94: 197–225.
Horowitz, Tamara & Gerald Massey (eds.) 1991. *Thought Experiments in Science and Philosophy*. Lanham: Rowman & Littlefield.
Horvath, Christopher 2000. "Interactionism and innateness in the evolutionary study of human nature", *Biology and Philosophy* 15: 321–337.
Hohwy, Jakob & Jesper Kallestrup (eds.) 2008. *Being Reduced. New Essays on Reduction, Explanation, and Causation*. Oxford: Oxford University Press.
Hruschka, Daniel J. & Joseph Henrich 2006. "Friendship, Cliquishness, and the Emergence of Cooperation". *Journal of Theoretical Biology* 239: 1–15.
Hruschka, Daniel J., Daniel H. Lende & Carol M. Worthman 2005. "Biocultural dialogues: Biology and culture in Psychological Anthropology". *Ethos* 33: 1–19.

Hull, David 1980. "Individuality and Selection", *Annual Review of Ecology and Systematics* 11: 311–332.
Hull, David 1981. "Units of evolution: a metaphysical essay", in Uffe Jensen & Rom Harré (eds.): *The philosophy of evolution*. Brighton: Harvester Press, 23–44.
Hull, David 1988a. "Interactors versus vehicles", in Henry Plotkin (ed.): *The Role of Behavior in Evolution*. Cambridge, MA: MIT Press, 19–50.
Hull, David 1988b. *Science as a Process: An evolutionary account of the social and conceptual development of science*. Chicago: The University of Chicago Press.
Huneman, Philippe (ed.) 2013. *Functions: selection and mechanism*. Dordrecht: Springer.
Hunt, Gavin R. & Russell D. Gray 2003. "Diversification and cumulative evolution in New Caledonian crow tool manufacture", *Proceedings of Royal Society of London B* 270: 867–874.
Hurlburt, Russell & Eric Schwitzgebel 2007. *Describing Inner Experience? Proponent Meets Skeptic*. Cambridge, MA: MIT Press.
Hurley, Susan & Matthew Nudds 2006. "The questions of animal rationality: Theory and evidence", In Susan Hurley & Nudds (eds.): *Rational animals?* Oxford, England: Oxford University Press, 1–83.
Hutto, Daniel D. 2007. "Folk Psychology without Theory or Simulation", in Daniel Hutto & Matthew Ratcliffe (eds.): *Folk Psychology Reassessed*. Springer, 115–135.
Hutto, Daniel D. 2008. *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. MIT press.
Hutto, Daniel D. 2009. "Folk Psychology as Narrative Practice", *Journal of Consciousness Studies* 16: 9–39.
Hutto, Daniel D., Mitchell Herschbach & Victoria Southgate 2011. "Editorial: Social Cognition: Mindreading and Alternatives", *Review of Philosophy and Psychology* 2.
Hutto, Daniel D. & Erik Myin 2013. *Radicalizing Enactivism: Basic Minds Without Content*. Cambridge: MIT Press.
Hutto, Daniel D. & Erik Myin 2017. *Evolving Enactivism – Basic Minds Meet Content*. Cambridge: MIT Press.
Hutto, Daniel D. & Matthew Ratcliffe (eds.) 2007. *Folk Psychology Reassessed*. Springer.
Huxley, Julian 1942. *Evolution – the Modern Synthesis*. London: Allen & Unwin.
Hyland, Michael E. 1994. "Different Vehicles for Group Selection in Humans", *Behavioral and Brain Sciences* 17: 628.
Ignacio Galparsoro, José & Alberto Cordero (eds.) 2013a. *Reflections on Naturalism*. Rotterdam: Sense Publishers.

Ignacio Galparsoro, José & Alberto Cordero 2013b. "Introduction: Naturalism and Philosophy", in José Ignacio Galparsoro & Alberto Cordero (eds.): *Reflections on Naturalism*. Rotterdam: Sense Publishers, 1–15.

Ingold, Tim (ed.) 1995a. *Companion Encyclopedia of Anthropology*. London: Routledge.

Ingold, Tim 1995b. "General Introduction." In Tim Ingold (ed.): *Companion Encyclopedia of Anthropology*. London: Routledge. xiii–xxii.

Ingold, Tim 2001. "From complementarity to obviation: on dissolving the boundaries between social and biological anthropology, archeology, and psychology". In Susan Oyama, Paul E. Griffiths & Russell D. Gray (eds.): *Cycles of Contingency: Developmental Systems and Evolution*. Cambridge, Ma: MIT Press.

Ingold, Tim 2007. "The trouble with 'evolutionary biology'", *Anthropology Today* 23: 13–17.

Jablonka, Eva & Marion J. Lamb 1995. *Epigenetic Inheritance and Evolution: the Lamarckian Dimension*. Oxford: Oxford University Press.

Jablonka, Eva & Marion J. Lamb 2005. *Evolution in Four Dimensions*. Cambridge: MIT Press.

Jablonka, Eva & Marion J. Lamb 2008. "Soft inheritance: challenging the modern synthesis", *Genetics and Molecular Biology* 31: 389–395.

Jackson, Frank 1998. *From Metaphysics to Ethics*. Oxford: Clarendon Press.

Jacobs, Armand & Odile Petit 2011. "Social Network Modeling: A Powerful Tool for the Study of Group Scale Phenomena in Primates", *American Journal of Primatology* 73: 741–747.

James, William 1884. "On Some Omissions of Introspective Psychology", *Mind* 33: 1–11.

Jensen, Uffe, & Rom Harré (eds.) 1981. *The philosophy of evolution*. Brighton: Harvester Press.

Johnson, Dominic D. P., Pavel Stopka & Josh Bell 2002. "Individual variation evades the Prisoner's Dilemma", *BMC Evolutionary Biology* 2. URL: http://www.biomedcentral.com/1471-2148/2/15.

Jones, Steve, Robert Martin, & David Pilbeam (eds.) 1992. *The Cambridge Encyclopedia of Human Evolution*. Cambridge: Cambridge University Press.

Joyce, Richard 1998. *The Myth of Morality*. Cambridge, Ma: Cambridge University Press.

Joyce, Richard 2006. *Evolution of Morality*. Cambridge, Ma: MIT Press.

Kahneman, Daniel 2003. "Maps of bounded rationality: Psychology for behavioral economics", *The American economic review* 93: 1449–1475.

Kahneman, Daniel 2011. *Thinking, fast and slow*. Macmillan.

Kahneman, Daniel, Paul Slovic & Amos Tversky (eds.) 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.
Kahneman, Daniel & Amos Tversky 1979. "Prospect Theory: An Analysis of Decision under Risk", *Econometrica* 47: 263–91.
Kane, Robert (ed.) 2002. *Oxford Handbook on Free Will*. New York: Oxford University Press.
Kant, Immanuel 1785. *Grundlegung zur Metaphysik der Sitten*. (In English: *Groundwork of the metaphysics of morals.* Translated by Mary J. Gregor. Cambridge: Cambridge University Press, 1998.)
Kant, Immanuel 1790. *Kritik der Urteilskraft.* (In English: *Critique of Judgement*. Translated by James Creed Meredith. Oxford: Oxford University Press, 2007 [1952].)
Kaplan, Hillard, Kim Hill, Jane Lancaster & A. Magdalena Hurtado 2000. "A Theory of Human Life History Evolution: Diet, Intelligence, and Longevity", *Evolutionary Anthropology* 9: 156–185.
Karmiloff-Smith, Annette 1992. *Beyond Modularity: A Developmental Perspective to Cognitive Science*. Cambridge: MIT Press.
Karmiloff-Smith, Annette 1998. "Development itself is the key to understanding developmental disorders", *Trends in Cognitive Sciences* 2: 389–398.
Karmiloff-Smith, Annette 2006. "Ontogeny, genetics, and evolution: a perspective from developmental neuroscience", *Biological Theory* 1: 44–51.
Katz, Leonard D. (ed.) 2000. *Evolutionary Origins of Morality: Cross-Disciplinary Perspectives*. Imprint Academic, Thorverton. (*Journal of Consciousness Studies* 7).
Katz, Paul S. 2011. "Neural mechanisms underlying the evolvability of behavior", *Philosophical Transactions of Royal Society of London, B: Biological Sciences*, 366: 2086–2099.
Keil, Frank C. 1989. *Concepts, Kinds and Cognitive Development*. Cambridge, Mass.: Bradford Books/MIT Press.
Keil, Frank C. 2003. "Folkscience: coarse interpretations of a complex reality", *Trends in Cognitive Sciences* 7: 368–373.
Keller, Evelyn Fox & Elizabeth Lloyd (ed.) 1992. *Keywords in Evolutionary Biology*. Harvard University Press, Cambridge, MA.
Keller, L. & K. G. Ross 1998. "Selfish genes: a green beard in the red fire ant", *Nature* 394: 573–575.
Kerr, Benjamin & Peter Godfrey-Smith 2002. "Individualist and multi-level perspectives on selection in structured populations", *Biology and Philosophy* 17: 477–517.
Kerr, Benjamin, Peter Godfrey-Smith & Marcus W. Feldman 2004. "What is altruism?" *Trends in Ecology and Evolution* 19: 135–140.

Keysara, Boaz, Shuhong Lin & Dale J. Barr 2003. "Limits on theory of mind use in adults", *Cognition* 89: 25–41.
Khalidi, Muhammad Ali 2002. "Nature and Nurture in Cognition", *British Journal for the Philosophy of Science* 53: 251–272.
Khalidi, Muhammad Ali 2007. "Innate cognitive capacities", *Mind & Language* 22: 92–115.
Kiikeri, Mika & Tomi Kokkonen 2007. "Biological Notions of Innateness and Explanation of Language Acquisition." In Johannes Persson & Petri Ylikoski (eds.): *Rethinking Explanation*. Dordrecht: Springer, 177–192.
Kim, Jaegwon 1989. "The myth of nonreductive physicalism". Reprinted 1993 in Jaegwon Kim (ed.): *Supervenience and mind*. Cambridge: Cambridge University Press, 265–284.
Kim, Jaegwon 1992. "'Downward causation' in emergentism and nonreductive physicalism", in Ansgar Beckermann, Hans Flohr & Jaegwon Kim (eds.): *Emergence or reduction?* Berlin: Walter de Gruyter, 119–138.
Kim, Jaegwon 1993. *Mind and Supervenience*. Cambridge: Cambridge University Press.
Kim, Jaegwon 2005. *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.
Kinzey, Warren G. (ed.) 1987. *The Evolution of Human Behavior: Primate Models*. Albany, NY: SUNY Press.
Kitayama, Shinobu & Jiyoung Park 2010. "Cultural neuroscience of the self: understanding the social grounding of the brain". *Social, Connective and Affective Neuroscience* 5: 111–129.
Kitcher, Philip 1984. "1953 and all that: A tale of two sciences". *Philosophical Review* 93: 335–373.
Kitcher, Philip 1985. *Vaulting Ambition: Sociobiology and the Quest for Human Nature*. MIT Press, Cambridge.
Kitcher, Philip 1989. "Explanatory Unification and the Causal Structure of the World", in Philip Kitcher & Wesley Salmon (eds.): *Scientific Explanation*. Minneapolis: University of Minnesota Press, 410–505.
Kitcher, Philip 1993a. "The evolution of human altruism", *Journal of Philosophy* 90: 148–167.
Kitcher, Philip 1993b. "Function and design", in Peter French et al (eds.): *Midwest Studies in Philosophy* xviii.
Kitcher, Philip 1993c. *The Advancement of Science*. Oxford: Oxford University Press.
Kitcher, Philip 1997. "Psychological altruism, evolutionary origins and moral rules", *Philosophical Studies* 80: 283–316.

Kitcher, Philip 2014. *The Ethical Project*. Cambridge, Ma: Harvard University Press.
Klaczynsky, Paul A. 2009. "Cognitive and social development: Dual-process research and theory", in Jonathan St.B.T. Evans & Keith Frankish (eds.), *In Two Minds: Dual Processes and Beyond*. Oxford University Press, 265–292.
Klopher, Peter H. 2001. "Paternal Care and Development". In Susan Oyama, Paul E. Griffiths & Russell D. Gray (eds.): Cycles of Contingency: Developmental Systems and Evolution. Cambridge, Ma: MIT Press. 167–174.
Knobe, Joshua 2007. "Folk psychology: science and morals", in Daniel D. Hutto & Matthew Ratcliffe (eds.) (2007): *Folk Psychology Reassessed*. Dordrecht: Springer, 157–173.
Knobe, Joshua, & Shaun Nichols (eds.) 2008. *Experimental Philosophy*. New York: Oxford University Press.
Knobe, Joshua, Tania Lombrozo, & Shaun Nichols (eds.) 2014. *Oxford Studies in Experimental Philosophy*, volume 1. Oxford: Oxford University Press.
Kokkonen, Tomi 2003. "Evoluutio ja altruismi", *Ajatus* 60: 241–286.
Kokkonen, Tomi 2011. "Mielen teoria, selittäminen ja ymmärtäminen", *Tiede & edistys* 4/2011: 277–290.
Kokkonen, Tomi 2012. "Arkipsykologia oman ja muiden toiminnan ymmärtämisessä", in Valtteri Viljanen, Helena Siipi & Matti Sintonen (eds.): *Ymmärrys*. Reports form the Department of Philosophy. Turku: University of Turku, 319–329.
Kokkonen, Tomi 2016. "Kuinka mahdollisesti selittää? Historialliset selitykset ja evidenssi." In Ilkka Niiniluoto, Tuomas Tahko & Teemu Toppinen (eds.): *Mahdollisuus*. Helsinki: Suomen Filosofinen Yhdistys.
Kokkonen, Tomi & Samuli Pöyhönen 2012. "Olemukset piilopremisseinä argumentaatiossa", in Juho Ritola (ed.): *Tutkimuksia argumentaatiosta*. Turku: Uniprint, 191–206.
Kornblith, H. 1994. "What is naturalistic epistemology?" In Kornblith (ed.): *Naturalizing epistemology*. 2nd edition. Cambridge, MA: MIT Press. 1–14.
Koskinen, Inkeri 2014. "At least two concepts of culture", in *Folklore* 125: 267–285.
Krebs, John & Nick Davies (eds.) 1997. *Behavioural Ecology: An Evolutionary Approach*. Fourth Edition. Blackwell.
Krickel, Beate 2018. *The Mechanical World: The Metaphysical Commitments of the New Mechanistic Approach*. Dordrecht: Springer.
Kronfeldner, Maria 2019. *What's Left of Human Nature*. Cambridge, Ma: MIT Press.
Kronfeldner, Maria 2021. "Digging the channels of inheritance: On how to distinguish between cultural and biological inheritance", *Philosophical Transactions of Royal Society*, published online 17 May 2021, https://doi.org/10.1098/rstb.2020.0042

Kropotkin, Peter 1904. *Mutal Aid: A Factor of Evolution*. London: William Heinemann.
Kuorikoski, Jaakko 2009. "Two concepts of mechanism: componential causal system and abstract form of interaction". *International Studies in the Philosophy of Science* 23: 143–160.
Kuorikoski, Jaakko 2012. "Mechanisms, Modularity and Constitutive Explanation", *Erkenntnis* 77: 361–380.
Kuorikoski, Jaakko & Petri Ylikoski 2008. "Intentional Fundamentalism", in Alexander Hieke & Hannes Leitgeb (eds.): *Reduction and Elimination in Philosophy and the Sciences - Papers of the 31st International Wittgenstein Symposium Vol XVI*. Kirchberg am Wechsel, Austria: Austrian Ludwig Wittgenstein Society.
Kuper, Adam 1999. *Culture: The Anthropologists' Account*. Cambridge: Harvard University Press.
Lakatos, Imre 1970. "Falsification and the methodology of scientific research programmes", in Imre Lakatos & Alan Musgrave (eds.): *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press, 91–195.
Lakatos, Imre 1978. *The Methodology of Scientific Research Programmes: Philosophical Papers Volume 1*. Cambridge University Press, Cambridge.
Lakatos, Imre & Alan Musgrave (eds.) 1970. *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press.
Laland, Kevin N., & Gillian Brown 2002. *Sense and Nonsense: Evolutionary Perspectives on Human Behaviour*. Oxford: Oxford University Press.
Laland, Kevin N. & Bennet G. Galef 2009. *The Question of Animal Culture*. Harvard University Press, Cambridge
Laland, Kevin N., John Odling-Smee & Marcus W. Feldman 2000a. "Niche construction, biological evolution, and cultural change", *Behavioral and Brain Sciences* 23: 131–175.
Laland, Kevin N., John Odling-Smee & Marcus W. Feldman 2000b. "Group Selection: A Niche Construction Perspective", *Journal of Consciousness Studies* 7: 221–225.
Laland, Kevin N., John Odling-Smee & Marcus W. Feldman 2001. "Cultural niche construction and human evolution", *Journal of Evolutionary Biology* 14: 22–33.
Laland, Kevin N., Kim Sterelny, John Odling-Smee, William Hoppitt & Tobias Uller 2011. "Cause and effect in biology revisited: is Mayr's proximate–ultimate dichotomy still useful?" *Science* 334: 1512–1516.
Laland, Kevin N., John Odling-Smee, William Hoppitt & Tobias Uller 2013. "More on how and why: cause and effect in biology revisited", *Biology & Philsoophy* 28: 719–745.

Laland, Kevin N. & Michael J. O'Brien 2012. "Cultural Niche Construction: An Introduction", *Biological Theory* 6, 191–202.

Laland, Kevin N., Tobias Uller, Marc Feldman, Kim Sterelny, Gerd B. Müller, Armin Moczek, Eva Jablonka, John Odling-Smee, Gregory A. Wray, Hopi E. Hoekstra, Douglas J. Futuyma, Richard E. Lenski, Trudy F. C. Mackay, Dolph Schluter & Joan E. Strassmann 2014. "Does evolutionary theory need a rethink?" *Nature* 514 (08 October 2014).

Laland, Kevin N., Tobias Uller, Marc Feldman, Kim Sterelny, Gerd B. Müller, Armin Moczek, Eva Jablonka & John Odling-Smee 2015. "The extended evolutionary synthesis: its structure, assumptions and predictions", *Proceedings of the Royal Society B* 282: 20151019. http://dx.doi.org/10.1098/rspb.2015.1019

Leaky, Richard & Roger Lewin 1992. *Origins Reconsidered: In Search of What Makes Us Human*. New York: Doubleday.

Lehrman, Daniel 1953. "A critique of Konrad Lorenz's theory of instinctive behavior", *The Quarterly Review of Biology* 28: 337–363.

Leigh, Egbert G., Jr. 2010a. "The group selection controversy", *The Journal of Evolutionary Biology* 23: 6–19.

Leigh, Egbert G., Jr. 2010b. "The evolution of mutualism", *The Journal of Evolutionary Biology* 23: 2507–2528.

Levinson, Stephen, & Pierre Jaisson (eds.) (2006): *Evolution and Culture*. Cambridge, Ma: The MIT Press.

Levitis, Daniel, William Lidicker & Glenn Freund (2009): "Behavioural Biologists Do Not Agree on What Constitutes Behaviour", *Animal Behaviour* 78: 103–110.

Levy, Arnon 2013. "Three Kinds of "New Mechanism", *Biology & Philosophy* 28: 99–114.

Lewens, Tim 2000. "Function Talk and the Artefact Model", *Studies in History and Philosophy of Biology and Biomedical Sciences* 31: 95–111.

Lewens, Tim 2002. "Adaptationism and Engineering", *Biology and Philosophy* 17: 1–31.

Lewens, Tim 2004. *Organisms and Artifacts: Design in Nature and Elsewhere*. Cambridge, MA: MIT Press.

Lewens, Tim 2009. "Seven Types of Adaptationism", *Biology and Philosophy* 24: 161–182.

Lewens, Tim 2010. "The Natures of Selection", *British Journal for the Philosophy of Science* 61: 313–333.

Lewens, Tim 2015. *Cultural Evolution: Conceptual Challenges*. Oxford: Oxford University Press.

Lewin, Roger, & Robert A. Foley (2004): *Principles of Human Evolution*. Second Edition. Oxford: Blackwell Publishing.
Lewis, David 1972. "Psychophysical and Theoretical Identifications," *Australasian Journal of Philosophy* 50: 249–58.
Lewis, David 1973. "Causation", *Journal of Philosophy* 70: 556–567.
Lewontin, Richard 1970. "The units of selection", *Annual Review of Ecology and Systematics* 1: 1–18.
Lewontin, Richard 1974. "The analysis of variance and the analysis of causes", *American Journal of Human Genetics* 26: 400–411.
Lewontin, Richard C. 1984. "Adaptation". In Elliot Sober (ed.): *Conceptual Issues in Evolutionary Biology*. The MIT Press, Cambridge, MA, 235–251.
Lieberman, Bruce S. & Elisabeth S. Vrba 2005. "Stephen Jay Gould on species selection: 30 years of insight", *Paleobiology* 31: 113–121.
Lieberman, Matthew 2007. "Social Cognitive Neuroscience: A Review of Core Processes", *Annual Review of Psychology* 58: 259–289.
Linquist, Stefan, Edouard Machery, Paul E. Griffiths & Karola Stotz 2001. "Exploring the folkbiological conception of human nature", *Philosophical Transactions of the Royal Society B* 366: 444–453.
Lion, Sébastien, Vincent A.A. Jansen & Troy Day 2011. "Evolution in structured populations: beyond the kin versus group debate", *Trends in Ecology and Evolution* 26: 193–201.
Lipton, Peter 2000. "Introduction: The Pull of Teleology", *Studies in History and Philosophy of Biology and Biomedical Sciences* 31: 1–10.
List, Christian & Philip Pettit 2011. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. New York: Oxford University Press.
Lloyd, Elisabeth 1986. "Evaluation of Evidence in Group Selection Debates", *Proceedings of the Philosophy of Science Association 1986*: 483–493.
Lloyd, Elisabeth 1988. *The Structure and Confirmation of Evolutionary Theory*. Princeton: Princeton University Press.
Lloyd, Elisabeth 1992. "Unit of Selection". In Evelyn Fox Keller & Elisabeth A. Lloyd (eds.): *Keywords in Evolutionary Biology*. Cambridge, MA: Harvard University Press. 334–340.
Lloyd, Elisabeth 2001. "Units and Levels of Selection: An Anatomy of the Units of Selection Debates". In Rama S. Singh, Costas B. Krimbas, Diane B. Paul, and John Beatty (eds.): *Thinking About Evolution (Historical, Philosophical, and Political Perspectives, vol. 2)*. Cambridge: Cambridge University Press.
Lloyd, Elisabeth 2005. "Why the Gene Will Not Return", *Philosophy of Science* 72: 287–310.

Lloyd, Elisabeth 2006. *The Case of the Female Orgasm: Bias in the Science of Evolution*. Cambridge, MA: Harvard University Press.
Lloyd, Elisabeth 2015. "Adaptationism and the Logic of Research Questions: How to Think Clearly about Evolutionary Causes", *Biological Theory* 10: 343–362.
Lloyd, Elisabeth 2017 [2005]. "Units and levels of selection". In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition). https://plato.stanford.edu/archives/spr2020/entries/selection-units/. Original version 2005, substantive content change 2017, last modified 2020.
Lloyd, Elisabeth & Stephen J. Gould. 1993. "Species selection on variability", *Proceedings of the National Academy of Sciences USA* 90: 595–599.
Longino, Helen 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton: Princeton University Press.
Longino, Helen 2002. *The Fate of Knowledge*. Princeton: Princeton University Press.
Longino, Helen 2013. *Studying Human Behavior: How Scientists Investigate Aggression and Sexuality*. Chicago: University of Chicago Press.
Lorenz, Konrad 1965. *Evolution and Modification of Behavior*. Chicago: University of Chicago Press.
Lorenz, Konrad 1966 [1963]. *On Aggression*. Translated by Marjorie Latzke from German original *Das sogenannte Böse zur Naturgeschichte der Aggression*. London: Methuen Publishing.
Lorini, Giuseppe & Wojciech Żełaniec 2018. "The Background of Constitutive Rules", *Argumenta* 4: 9–19.
Love, Alan C. 2010. "Idealization in evolutionary developmental investigation: A tension between phenotypic plasticity and normal stages". *Philosophical Transactions of the Royal Society B: Biological Sciences* 365: 679–690.
Love, Alan C. 2014. "The erotetic organization of development". In Alessandro Minelli & Thomas Pradeu (eds.): *Towards a Theory of Development*. Oxford: Oxford University Press. 33–55.
Ludwig, Kirk 2016. *From Individual to Plural Agency: Collective Action I*. Oxford: Oxford University Press.
Ludwig, Kirk 2017. *From Plural to Institutional Agency: Collective Action II*. Oxford: Oxford University Press.
Lumsden, Charles J. & Edward O. Wilson 1981. *Genes, Mind and Culture: The Coevolutionary Process*. Cambridge, MA: Harvard University Press.
Luque, Victor J. 2016. "The Principle of Stasis: Why drift is not a Zero-Cause Law", *Studies in History and Philosophy of Biological and Biomedical Sciences* 57: 71–79.

Lurz, Robert (ed.) 2009. *The Philosophy of Animal Minds*. New York: Cambridge University Press.

Lyons, Derek E., Andrew G. Young & Frank C. Keil 2007. "The hidden structure of overimitation". *Proceedings of the National Academy of Sciences* 104: 19751–19756.

MacDonald, Kevin 1994. "Group Evolutionary Strategies: Dimensions and Mechanisms", *Behavioral and Brain Sciences* 17: 629–630.

Machamer, Peter 2004. "Activities and causation: the metaphysics and epistemology of mechanisms", *International Studies in the Philosophy of Science* 18: 27–39.

Machamer, Peter, Lindley Darden, & Carl F. Craver 2000. "Thinking about Mechanisms", *Philosophy of Science* 67: 1–24.

Machery, Edouard & Ron Mallon 2010. "Evolution of morality". In John Doris (ed.): *The Moral Psychology Handbook*. Oxford: Oxford University Press

Mackie, John L. 1974. *The Cement of the Universe: A Study of Causation*. Oxford University Press.

MacLagan, David S. 1932. "The effect of population density upon rate of reproduction with special reference to insects", *Proceedings of Royal Society B* 111: 437–454.

Mahner, Martin & Mario Bunge 2001. "Function and Functionalism: A Synthetic Perspective", *Philosophy of Science* 68: 75–94.

Malle, Bertram F. & Gale E. Pearce 2001. "Attention to behavioral events during interaction: Two actor–observer gaps and three attempts to close them", *Journal of Personality and Social Psychology* 81: 278–294.

Mallon, Ron & Stephen Stich 2000. "The Odd Couple: The Compatibility of Social Construction and Evolutionary Psychology", *Philosophy of Science* 67: 133–154.

Mallon, Ron & Jonathan Weinberg 2006. "Innateness as closed-process invariantism", *Philosophy of Science* 73, 323–344.

Mameli, Matteo 2001. "Mindreading, Mindshaping, and Evolution", *Philosophy and Biology* 16: 595–626.

Mameli, Matteo & Patrick Bateson 2006. "Innateness and the sciences", *Biology and Philosophy* 21, 155–188.

Mann, Michael 2004. *Fascists*. Cambridge: Cambridge University Press.

Manning, Aubrey 2005. "Four decades on from the 'four questions'", *Animal Biology* 55: 287–296.

Marcus, Ruth B. 1990. "Some revisionary proposals about belief and believing", *Philosophy and Phenomenological Research* 50: 132–153.

Marewski, Julian N., Wolfgang Gaissmaier & Gerd Gigerenzer 2010. "Good judgments do not require complex cognition", *Current Directions in Psychological Science* 11: 103–121.
Marshall, James A. 2011. "Group selection and kin selection: formally equivalent approaches", *Trends in Ecology and Evolution* 26: 325–332.
Martens, Johannes 2011. "Social evolution and strategic thinking", *Biology and Philosophy* 26: 697–715.
Marr, David 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.
Mathew, Sarah & Charles Perreault 2015. "Behavioural variation in 172 small-scale societies indicates that social learning is the main mode of human adaptation", *Proceedings of the Royal Society B* 282: 20150061. http://dx.doi.org/10.1098/rspb.2015.0061
Matthen, Mohan & Andrew Ariew 2002. "Two Ways of Thinking about Fitness and Natural Selection", *Journal of Philosophy* 99: 55–83.
Matthen, Mohan & Andrew Ariew 2009. "Selection and causation", *Philosophy of Science* 76: 201–224.
Maynard Smith, John 1964. "Group selection and kin selection", *Nature* 201: 1145–1147.
Maynard Smith, John 1982. *Evolution and the Theory of Games*. Cambridge University Press, Cambridge.
Maynard Smith, John 1988. "Evolutionary progress and the levels of selection" In Matthew H. Nitecki (ed.): *Evolutionary Progress*. Chicago: University of Chicago Press. 219–230.
Maynard Smith, John 1998. "The origin of altruism", *Nature* 393: 639–640.
Maynard Smith, John & George Price 1973. "The logic of animal conflict", *Nature* 246: 15–18.
Maynard Smith, John & Eors Szathmary 1997. *The Major Transitions in Evolution*. Oxford: Oxford University Press.
Mayr, Ernst 1959. "Typological vs. population thinking." In Betty Meggers (ed.): *Evolution and Anthropology: A Cenentennial Appraisal*, The Anthropological Society of Washington, Washingon, DC, 409–412.
Mayr, Ernst 1983. "How to carry out the adaptationist program?" *American Naturalist* 121: 324–334.
McAndrew, Francis T. 2002. "New evolutionary perspectives on altruism: multilevel-selection and costly signalling theories". In Alan E. Kazdin (ed.): *Current Directions in Psychological Science* 11: 79–82.

McBrearty, Sally, & Alison S. Brooks 2000. "The Revolution That Wasn't: A New Interpretation of the Origin on Modern Human Behavior", *Journal of Human Evolution* 39: 453–563.

McElreath, Richard & Robert Boyd 2007. *Mathematical Models of Social Evolution. A Guide for the Perplexd*. Chicago: Chicago University Press.

McGeer, Victoria 2007. "The Regulative Dimension of Folk Psychology." In Daniel D. Hutto & Matthew Ratcliffe (eds.): *Folk Psychology Re-Assessed*. Dordrecht: Springer, 137–156.

McGlothlin, Joel W., Allen J. Moore, Jason B. Wolf & Edmund D. Brodie III 2010. "Interacting phenotypes and the evolutionary process. III. Social evolution" *Evolution* 64: 2558–2574.

McGrew, William 1998. "Culture in non-human primates?" *Annual Review of Anthropolgy* 27: 301–328.

McGrew, William 2009. "Ten dispatches from the chimpanzee culture wars, plus postscript (revisiting the battlefronts)". In Kevin N. Laland & Bennett G. Galef (eds): *The question of animal culture*. Cambridge: Harvard University Press, 41–69.

McGuigan, Nicola & Murray Graham 2009. "Cultural transmission of irrelevant tool actions in diffusion chains of 3- and 5-year-old children". *European Journal of Developmental Psychology* 7. https://doi.org/10.1080/17405620902858125

McGuigan, Nicola, Jenny Makinson & Andrew Whiten 2011. "From over-imitation to super-copying: adults imitate causally irrelevant aspects of tool use with higher fidelity than young children." *British Journal of Psychology* 102: 1–18.

McGuigan, Nicola, Emily Burdett, Vanessa Burgess, Lewis Dean, Amanda Lucas, Gillian Vale & Andrew Whiten 2017. "Innovation and social transmission in experimental micro-societies: exploring the scope of cumulative culture in young children". *Philosophical Transactions of The Royal Society B: Biological Sciences* 372. https://doi.org/10.1098/rstb.2016.0425

Medin, Douglas & Scott Atran (eds.) 1999. *Folk Biology*. Cambridge MA: MIT Press.

Medin, Dougals & Scott Atran 2004. "The native mind: Biological categorization and reasoning in development and across cultures", *Psychological Review* 111: 960–983.

Medin, Douglas & Andrew Ortony 1989. "Psychological Essentialism". In Stella Vosniadou & Andrew Ortony (eds.): *Similarity and Analogical Reasoning*. Cambridge: Cambridge University Press.

Mele, Alfred 1992. *The Springs of Action*. New York: Oxford University Press.

Mele, Alfred 2000. "Goal-Directed Action: Teleological Explanations, Causal Theories and Deviance", *Philosophical Perspectives* 14: 279 – 300.

Mele, Alfred 2002. *Motivation and Agency*. Oxford: Oxford University Press.

Mele, Alfred 2009. *Effective Intentions: The power of conscious will*. Oxford: Oxford University Press.
Melis, Alicia 2018. "The evolutionary roots of prosociality: the case of instrumental helping". *Current Opinion in Psychology* 20: 82–86.
Menary, Richard 2010. "Introduction to the Issue on 4E Cognition", *Phenomenological Cognitive Science* 9: 459–463.
Menzies, Peter 2007. "Mental Causation on the Program Model." In G. Brennan, R. Goodin, F. Jackson, & M. Smith (eds.): *Common Minds: Themes from the Philosophy of Philip Pettit*. Oxford: Oxford University Press. 28–54.
Menzies, Peter 2012. "The causal structure of mechanisms", *Studies in History and Philosophy of Biological and Biomedical Sciences* 43: 796–805.
Menzies, Peter & Huw Price 1993. "Causation as a Secondary Quality", *British Journal for the Philosophy of Science* 44: 187–203.
Merton, Robert 1957. *Social Theory and Social Structure*. London: The Free Press of Glencoe.
Michel, George & Celia Moore 1995. *Developmental Psychobiology: An interdisciplinary science*. Cambridge, MA: MIT Press.
Michod, Richard 1997. "Cooperation and conflict in the evolution of individuality I. Multilevel selection of the organism", *American Naturalist* 149: 607–645.
Michod, Richard 1999. *Darwinian Dynamics*. Princeton, NJ: Princeton University Press.
Michod, Richard & Aurora Nedelcu 2003. "On the reorganization of fitness during evolutionary transitions in individuality", *Integrative and Comparative Biology* 43: 64–73.
Millikan, Ruth 1984. *Language, Thought and Other Biological Categories*. The MIT Press, Cambridge, MA.
Millikan, Ruth 2004. *Varieties of Meaning*. Cambridge, Mass: MIT Press.
Mills, Susan & John Beatty 1979. "The Propensity Interpretation of Fitness", *Philosophy of Science* 46: 263–286.
Millstein, Roberta L. 2006. "Natural Selection as a Population Level Causal Process", *British Journal for the Philosophy of Science* 57: 627–653.
Millstein, Roberta L., Robert A. Skipper & Michael R. Dietrich 2009. "(Mis)interpreting Mathematical Models: Drift as a Physical Process", *Philosophy and Theory in Biology* 1: e002.
Mitchell, Sandra 1992. "On Pluralism and Competition in Evolutionary Explanations", *American Zoologist* 32: 135–144.
Mitchell, Sandra 1997. "Pragmatic laws", *Philosophy of Science* 64: S486–S479.

Mitchell, Sandra 2000. "Dimensions of scientific laws", *Philosophy of Science* 67: 242–265.
Mitchell, Sandra 2002. "Integrative pluralism", *Biology and Philosophy* 17: 55–70.
Mitchell, Sandra 2003. *Biological Complexity and Pluralism*. Cambridge: Cambridge University Press.
Mithen, Steven 1995. "Understanding Mind and Culture: Evolutionary Psychology or Social Anthropology?" *Anthropology Today* 11: 3–7.
Mithen, Steven 1996. *The Prehistory of the Mind. The Cognitive Origins of Art and Science*. London: Thames & Hudson.
Mithen, Steven 2005. *The Singing Neanderthals: The Origin of Music, Language, Mind, and Body*. London: Weidenfeld & Nicolson.
Mitteldorf, Joshua & David Sloan Wilson 2000. "Population viscosity and the evolution of altruism", *Journal of Theoretical Biology* 204: 481–496.
Moll, Henrike & Michael Tomasello 2007. "Cooperation and human cognition: the Vygotskian intelligence hypothesis", *Philosophical Transactions of Royal Society of London B: Biological Sciences* 362: 639–648.
Moore, Allen J., Edmund D. Brodie III & Jason B. Wolf 1997. "Interacting phenotypes and the evolutionary process: I. Direct and indirect genetic effects of social interactions", *Evolution* 51: 1352–1362.
Morris, Douglas 2011. "Adaptation and habitat selection in the eco-evolutionary process", *Proceedings of the Royal Society B* 278: 2401–2411.
Moshman, David 2004. "From Inference to Reasoning: The Construction of Rationality", *Thinking & Reasoning* 10: 221–239.
Moss, Lenny 2001. "Deconstructing the Gene and Reconstructing Molecular Developmental Systems". In Susan Oyama, Paul E. Griffiths & Russell D. Gray (eds.): *Cycles of Contingency: Developmental Systems and Evolution*. Cambridge, Ma: MIT Press. 85–98.
Moss, Lenny 2004. *What Genes Can't Do*. Cambridge, Ma: The MIT Press.
Müller, Gerd 2007. "Evo-devo: extending the evolutionary synthesis", *Nature Review Genetics* 8: 943–949.
Müller, Gerd 2017. "Why an extended evolutionary synthesis is necessary". *Interface Focus* 7: 20170065. doi: 10.1098/rsfs.2017.0015
Murphy, Dominic 2006. *Psychiatry in the Scientific Image*. Cambridge, Ma: The MIT Press.
Nanay, Bence 2002. "The Return of the Replicator: What Is Philosophically Significant in a General Account of Replication and Selection?" *Biology and Philosophy* 17: 109–121.

Narvaez, Darcia 2014. *Neurobiology and the Development of Human Morality: Evolution, Culture and Wisdom*. New York: W.W. Norton.
Narvaez, Darcia, Jaak Panksepp, Allan Schore, & Tracy Gleason (eds.) 2012. *Evolution, Early Experience and Human Development: From Research to Practice and Policy*. New York: Oxford University Press.
Narvaez, Darcia, Kristin Valention, Agustin Fuentes, James J. McKenna & Peter Gray (eds.) 2014. *Ancestral Landscapes in Human Evolution: Culture, Childrearing and Social Wellbeing*. Oxford: Oxford University Press.
Narvaez, Darcia, Julia M. Braungart-Rieker, Laura E. Miller-Graff, Lee T. Gettler & Paul D. Hastings (eds.) 2016. *Contexts for Young Child Flourishing: Evolution, Family, and Society*.
Neander, Karen 1991. "Functions as Selected Effects: The Coceptual Analyst's Defence", *Philosophy of Science* 58: 168–184.
Newen, Albert, Leon De Bruin & Shaun Gallagher 2018. *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press.
Ney, Alyssa 2012. "Neo-positivist metaphysics", *Philosophical Studies* 160: 53–78.
Nichols, Shaun 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgment*. New York: Oxford University Press.
Nichols, Shaun & Stephen Stich 2003. *Mindreading. An Integrated Account of Pretence, Self-awareness, and Understanding Other Minds*. Oxford: Oxford University Press.
Nisbett, Richard & Lee Ross 1980. *Human Inference: Strategies and Shortcomings of Social Judgement*. Englewood Cliffs: Prentice Hall.
Nisbett, Richard & Timothy Wilson 1977. "Telling more than we can know: Verbal reports on mental processes", *Psychological Review* 84: 231–259.
Noë, Ronald 2006. "Cooperation experiments: Coordination through communication versus acting apart togerher". *Animal Behaviour* 7: 1–18.
Noë, Ronald & Peter Hammerstein 1994. "Biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating", *Behavioral Ecology and Sociobiology* 35: 1–11.
Noë, Ronald & Peter Hammerstein 1995. "Biological markets", *Trends in Ecology and Evolution* 10: 336–339.
Noë, Ronald & Peter Hammerstein 2016. "Biological trading and markets", *Philosophical Transactions of Royal Society of London B: Biological Sciences* 371: 20150101. http://dx.doi.org/10.1098/rstb.2015.0101
Nowak, Martin A., Corina E. Tarnita & Edward O. Wilson 2010. "The evolution of eusociality", *Nature* 466: 1057–1062.
Nunney, Leonard 2000. "Altruism, Benevolence and Culture", *Journal of Consciousness Studies* 7: 231–236.

O'Brien, Lilian 2019. "Action Explanation and its Presuppositions", *Canadian Journal of Philosophy* 49: 123–146.
Odling-Smee, John F., Kevin N. Laland & Marcus W. Feldman 2003. *Niche Construction: The Neglected Process in Evolution*. Princeton University Press.
Odling-Smee, John F. & Kevin N. Laland 2011. "Ecological Inheritance and Cultural Inheritance: What Are They and How Do They Differ?" *Biological Theory* 6: 220–230.
Ohtsuki, Hisashi & Yoh Iwasa 2004. "How should we define goodness? Reputation dynamics in indirect reciprocity". *Journal of Theoretical Biology* 231: 107–120.
Okasha, Samir 2002. "Genetic realtedness and the evolution of altruism", *Philosophy of Science* 69: 138–149.
Okasha, Samir 2005a. "Multilevel selection and the major transitions in evolution", *Philosophy of Science* 72: 1013–1025.
Okasha, Samir 2005b. "Maynard Smith on the levels of selection question", *Biology and Philosophy* 20: 989–1010.
Okasha, Samir 2006. *Evolution and the Levels of Selection*. Oxford University Press, Oxford.
Okasha, Samir 2016. "The relation between kin and multi-level Selection: An approach using causal graphs", *British Journal for the Philosophy of Science* 67: 435–470.
O'Neill 2015. "Relativizing innateness: innateness as the insensitivity of the appearance of a trait with respect to specified environmental variation". *Biology and Philosophy* 30: 211–225.
Oosterbeek, Hessel, Randolph Sloof & Gijs van de Kuilen 2004. "Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis", *Experimental Economics* 7: 171–188.
Oppenheim, Paul & Hilary Putnam 1958. "Unity of science as a working hypothesis". In H. Feigl, M. Scriven & G. Maxwell (eds.); *Concepts, theories, and the mind-body problem*. Minnesota studies in the philosophy of science II. Minneapolis, MN: University of Minnesota Press. 3–36.
Orzack, Steven H. & Elliot Sober 1994a. "Optimality models and the test of adaptationism", *The American Natutalist* 143: 361–380.
Orzack, Steven H. & Elliot Sober 1994b. "How (not) to test an optimality model", *Trends in Ecology and Evolution* 9: 265–267.
Orzack, Steven H. & Elliot Sober 1996. "How to formulate and test adaptationism", *The American Naturalist* 148: 202–210.
Orzack, Steven H. & Elliot Sober (eds.) 2001a. *Adaptationism and Optimality*. Cambridge: Cambridge University Press.

Orzack, Steven H. & Elliot Sober 2001b. "Adaptation, phylogenetic inertia, and the method of contorlled comparisons". In Steven H. Orzack & Elliot Sober (eds.): *Adaptationism and Optimality*. Cambridge: Cambridge University Press, 45–63.

Otsuka, Jun 2016. "A critical review of the statisticalist debate", *Biology and Philosophy* 31: 459–482.

Otsuka, Jun, Trin Turner, Colin Allen & Elisabeth A. Lloyd 2011. "Why the Causal View of Fitness Survives", *Philosophy of Science* 78: 209–224.

Overton, James A. 2011. "Mechanisms, Types, and Abstractions", *Philosophy of Science* 78: 941–954.

Oyama Susan 1985. *The Ontogeny of Information: Developmental Systems and Evolution*. Cambridge, Ma: Cambridge University Press.

Oyama, Susan, Paul E. Griffiths & Russell D. Gray 2001. *Cycles of Contingency. Developmental Systems and Evolution*. Cambridge, Mass.: MIT Press.

Padilla-Walker, Laura M., & Gustavo Carlo (eds.) 2014a. *Prosocial Development: A Multidimensional Approach*. Oxford: Oxford University Press.

Padilla-Walker, Laura M., & Gustavo Carlo 2014b. "The Study of Prosocial Behavior: Past, Present, and Future". In Laura M. Padilla-Walker & Gustavo Carlo (eds.): *Prosocial Development: A Multidimensional Approach*. Oxford: Oxford University Press.

Pagel, Mark 2012. *Wired for Culture: The Natural History of Human Cooperation*. Penguin.

Palmer, David C. 2006. "On Chomsky's Appraisal of Skinner's *Verbal Behavior:* A Half Century of Misunderstanding", *The Behavior Analyst* 29: 253–267.

Papineau, David 1984. "Representation and Explanation", *Philosophy of Science* 51: 550–72.

Paul, Ellen Frankel, Fred D. Miller, Jr., & Jeffrey Paul (eds.) 1993. *Altruism*. Cambridge University Press, Cambridge.

Paul, Robert 2018. "Culture from the perspective of dual inheritance", in Naomi Quinn (ed.): *Advances in Culture Theory from Psychological Anthropology*. London: Palgrave MacMillan, 47–74.

Paulus, Markus 2018. "The multidimensional nature of early prosocial behavior: a motivational perspective", *Current Opinion in Psychology* 20: 111–116.

Pearce, Trevor 2011. "Evolution and Constraints on Variation: Variant Specification and Range of Assessment", *Philosophy of Science* 78: 739–751.

Pearl, Judea 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, Raymond 1932. "The influence of density of population upon egg production in Drosophila Melanogaster", *Journal of Experimental Zoology* 63: 57–84.
Penn, Derek & Daniel Povinelli 2007. "On the lack of evidence that non-human animals possess anything remotely resembling a 'Theory of Mind'. *Philosophical Transactions: Biological Sciences* 362: 731–744.
Penrose, Roger 2016. *The Road to Reality: A Complete Guide to the Laws of the Universe*. Random House.
Pepper, John W. 2000. "Relatedness in Trait Group Models of Social Evolution", *Journal of Theoretical Biology* 206: 355–368.
Peressini, Anthony 1993. "Generalizing evolutionary altruism", *Philosophy of Science* 60: 568–586.
Perner, Josef & Anton Kühberger 2005. "Mental Simulation: Royal Road to Other Minds?" in Bertram F. Malle & Sara D. Hodges (eds.): *Other Minds: How Humans Bridge the Divide Between Self and Others*. New York: Guilford Press, 174–187.
Pernu, Tuomas K. 2013. "Does the interventionist notion of causation deliver us from the fear of epiphenomenalism?" *International Studies in the Philosophy of Science* 27: 157–172.
Persson, Johannes 2012. "Three Conceptions of Explaining How Possibly—and One Reductive Account". In Henk W. de Regt, Stephan Hartmann & Samir Okasha (eds.): EPSA Philosophy of Science: Amsterdam 2009. Springer. 275–286.
Persson, Johannes & Petri Ylikoski (eds.) 2007. *Rethinking Explanation*. Dordrecht: Springer.
Pettit, Philip 1993. *Common Mind: An Essay on Psychology, Society, and Politics*. Oxford: Oxford University Press.
Pigliucci, Massimo 2007. "Do we need an extended evolutionary synthesis?", *Evolution* 61: 2743–2749.
Pigliucci, Massimo 2008. "Is evolvability evolvable?" *Nature Reviews: Genetics* 9: 75–82.
Pigliucci, Massimo & Gerd B. Müller (eds.) 2010. *Evolution – the Extended Synthesis*. Cambridge, Mass.: MIT Press.
Pigliucci, Massimo, Courtney J. Murren & Carl D. Schlichting 2006. "Phenotypic plasticity and evolution by genetic assimilation", *The Journal of Experimental Biology* 209: 2362–2367.
Pinker, Steven 1998. *How the Mind Works*. Norton, New York.
Pinker, Steven 2002. *The Blank Slate: The Modern Denial of Human Nature*. Viking, New York
Pitt, Joseph C. 2001. "The dilemma of case studies: toward a Heraclitian philosophy of science", *Perspectives on Science* 9: 373–382.

Plotkin, Henry (ed.) 1988. *The Role of Behavior in Evolution*. Cambridge, MA: MIT Press.
Plotkin, Henry 1997. *Evolution in Mind: An Introduction to Evolutionary Psychology*. Harvard University Press, Cambridge, MA.
Pollock, G., & Lee Alan Dugatkin 1992. "Reciprocity and the emergence of reputation", *Journal of Theoretical Biology* 159: 25–37.
Poon, Connie S. K., Derek J. Koehler & Roger Buehler 2014. "On the psychology of self-prediction: Consideration of situational barriers to intended actions", *Judgment and Decision Making* 9: 207–225.
Portin, Peter 1993. "The concept of the gene: short history and present status", *The Quarterly Review of Biology* 68: 173–223.
Potochnik, Angela 2007. "Optimality modeling and explanatory generality", *Philosophy of Science* 74: 680–691.
Potochnik, Angela 2010. "Explanatory independence and epistemic interdependence: a case study of the optimality approach", *British Journal for the Philosophy of Science* 61: 213–233.
Potochnik, Angela & Brian McGill 2012. "The limitations of hierarchical organization". *Philosophy of Science* 79: 120–140.
Povinelli, Daniel J. & Jennifer Vonk 2003. "Chimpanzee minds: suspiciously human?" *Trends in Cognitive Sciences* 7: 157–160.
Prentice, Deborah & Dale Miller 2007. "Psychological essentialism of human categories", *Current Directions in Psychological Science* 16: 202–206.
Preston, Stephanie 2007. "A perception-action model for empathy", in Tom Farrow & Peter Woodruff (eds.): *Empathy in Mental Illness*. Cambridge: Cambridge University Press.
Preston, Stephanie & Frans de Waal 2002. "Empathy: its ultimate and proximate bases", *Behavioral and Brain Sciences* 25: 1–71.
Price, George 1970. "Selection and covariance", *Nature* 227: 520–521.
Price, Huw & Richard Corry (eds.) 2007. *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford University Press.
Prinz, Jess 2004. *Gut Reactions: A Perceptual Theory of Emotion*. New York: Oxford University Press.
Pruitt, Jonathan N. & Charles J. Goodnight 2014. "Site-specific group selection drives locally adapted group compositions", *Nature* 514: 359–362.
Pruitt, Jonathan N. & Charles J. Goodnight 2015. "Pruitt & Goodnight reply", *Nature* 524: doi:10.1038/nature14597.
Pullum, Geoffrey K. & Barbara C. Scholz 2002. "Empirical assessment of stimulus poverty arguments", *The Linguistic Review* 19: 9–50.

Putnam, Hilary 1967. "The 'innateness hypothesis' and explanatory models in linguistics," *Synthese* 17: 12–22.
Putnam, Hilary 1975. *Mind, Language, and Reality*. Cambridge: Cambridge University Press.
Quine, Willard Van Orman 1951. "Two Dogmas of Empiricism", *The Philosophical Review* 60: 20–43.
Quine, Willard Van Orman 1953. *From a Logical Point of View*. Cambridge, Ma: Harvard University Press.
Quinn, Naomi (ed.) 2018. *Advances in Culture Theory from Psychological Anthropology*. London: Palgrave MacMillan.
Raatikainen, Panu 2010. "Causation, exclusion, and the special sciences", *Erkenntnis* 73: 349–363.
Raerinne, Jani 2011. *Generalizations and Models in Ecology: Lawlikeness, Invariance, Stability, and Robustness*. Academic Dissertation, University of Helsinki. http://urn.fi/URN:ISBN:978-952-10-6768-6
Raff, Rudolf 1996. *The Shape of Life*. Chicago, IL: University of Chicago Press.
Railton, Peter 1978. "A Deductive-Nomological Model of Probabilistic Explanation", *Philosophy of Science* 45: 206–226.
Railton, Peter 1981. "Probability, Explanation, and Information", *Synthese* 48: 233–256.
Ramsey, Grant 2013a. "Organisms, Traits, and Population Subdivisions: Two Arguments against the Causal Conception of Fitness?" *British Journal for the Philosophy of Science*, online first, doi: 10.1093/bjps/axs010.
Ramsey, Grant 2013b. "Culture in humans and other animals", *Biology and Philosophy* 28: 457–479.
Rapoport, Anatol 1994. "Nongenetic and Non-Darwinian Evolution", *Behavioral and Brain Sciences* 17: 634.
Ratcliffe, Matthew 2000. "The Function of Function", *Studies in History and Philosophy of Biology and Biomedical Sciences* 31: 113–133.
Rawls, John 1955. "Two concepts of rules". *The Philosophical Review* 64: 3–32.
Reeve, Hudson & Paul Sherman 1993. "Adaptation and the goals of evolutionary research", *Quarterly Review of Biology* 68: 1–32.
Reeve, Hudson & Paul Sherman 2001. "Optimality and phylogeny: a critique of current thought". In Steven H. Orzack & Elliot Sober (eds.): *Adaptationism and Optimality*. Cambridge: Cambridge University Press, 64–113.
Reichenbach, Hans 1949. "The philosophical significance of the theory of relativity", in P. A. Schilpp (ed.): *Albert Einstein: Philosopher-scientist*. La Salle, IL: Open Court, 287–311.

Reisman, Kenneth &d Patrick Forber 2005. "Manipulation and the Causes of Evolution", *Philosophy of Science* 72: 1113–1123.

Resnik, David B. 1991. "How-possibly explanations in biology." *Acta Biotheoretica* 39: 141–149.

Reydon, Thomas 2012. "How-possibly explanations as genuine explanations and helpful heuristics: A comment on Forber". *Studies in History and Philosophy of Biological and Biomedical Sciences* 43: 302–310.

Richards, Robert 1987. *Darwin and the Emergence of Evolutionary Theories of Mind and Behavior*. University of Chicago Press, Chicago.

Richardson, Michael K. & Ariel D. Chipman 2003. "Developmental constraints in a comparative framework: a test case using cariations in pahalnx number during amniote evolution", *Journal of Experimental Zoology* 296B: 8–22.

Richardson, Robert 2007. *Evolutionary Psychology as Maladapted Psychology*. Cambridge: MIT Press.

Richerson, Peter & Robert Boyd 2005. *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago: Chicago University Press.

Richerson, Peter & Joseph Henrich 2012. "Tribal social instincts and the cultural evolution of institutions to solve collective action problems", *Cliodynamics* 3: 38–80.

Risjord, Mark 2000. *Woodcutters and Witchcraft*. Albany: SUNY Press.

Risjord, Mark 2005. "Reasons, causes, and action explanations", *Philosophy of the Social Sciences* 35: 1-13

Robbins, Philip & Murat Aydede (eds.) 2008. *The Cambridge Handbook of Situated Cognition*.

Robins, Sarah, John Symons & Paco Calvo (eds.) 2009. *The Routledge Companion to Philosophy of Psychology*. Routledge.

Roeboeks, Wil (ed.) 2007. *Guts, Brains, Food, and the Social Life of Early Hominids*. Leiden: University of Leiden Press.

Romeo, Rachel R., Julia A. Leonard, Sydney T. Robinson, Martin R. West, Allyson P. Mackey, Meredith L. Rowe & John D. E. Gabrieli 2018. "Beyond the 30-Million-Word Gap: Children's Conversational Exposure Is Associated with Language-Related Brain Function", *Psychological Science* 2: 700-710.

Rosas, Alejandro 2002. "Psychological and Evolutionary Evidence for Altruism", *Biology and Philosophy* 17: 93–107.

Rose, Eva, Peter Nagel & Daniel Haag-Wackernagel 2006. "Spatio-temporal use of the urban habitat by feral pigeons (*Columba livia*)", *Behavioral Ecology and Sociobiology* 60: 242–254.

Rose, Steven, Richard Lewontin & Leon Kamin 1984. *Not in Our Genes: Biology, Ideology, and Human Nature*. London: Penguin Press.
Rosenberg, Alexander 1978. "The Supervenience of Biological Concepts", *Philosophy of Science* 45: 368–386.
Rosenberg, Alexander 1992. "Altruism: theoretical contexts", in Evelyn Fox Keller & Elisabeth Lloyd (eds.): *Keywords in Evolutionary Biology*, Cambridge, MA: Harvard University Press, 19–28.
Ross, Don 2007. *Economic Theory and Cognitive Science: Microexplanation*. A Bradford Book.
Ross, Don 2011. "Estranged parents and a schizophrenic child: Choice in economics, psychology and neuroeconomics", *Journal of Economic Methodology* 18: 217–231.
Ross, Don 2013. "The evolution of individualistic norms." In Kim Sterelny, Richard Joyce, Brett Calcott & Ben Fraser (eds.): *Cooperation and Its Evolution*. Cambridge, Ma: The MIT Press, 17–44.
Roughley, Neil & Kurt Bayertz (eds.) 2019. *The Normative Animal? On the Anthropological Significance of Social, Moral, and Linguistic Norms*. Oxford: Oxford University Press.
Ruse, Michael 1998 [1986]. *Taking Darwin Seriously*, second edition, Prometheus Books, Amherst, New York.
Ruse, Michael & Edward O. Wilson 1986. "Moral Philosophy as Applied Science", *Philosophy* 61: 173–192.
Ryan, Michael 2005. "The evolution of behaviour, and integrating it towards a complete and correct understanding of behavioural biology", *Animal Biology* 55: 419–439.
Ryle, Gilbert 1949. *The Concept of Mind*. Chicago: University of Chicago Press.
Salmon, Wesley 1971. "Statistical Explanation", in Wesley Salmon (ed.): *Statistical Explanation and Statistical Relevance*. Pittsburgh: University of Pittsburgh Press, 29–87.
Salmon, Wesley 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.
Samuels, Richard 2002. "Nativism in Cognitive Science", *Mind and Language* 17, 233–265.
Samuels, Richard 2004. "Innateness in Cognitive Science", *Trends in Cognitive Sciences* 8, 136–141.
Samuels, Richard 2005. "The Complexity of Cognition: Tractability Arguments for Massive Modularity", in Peter Carruthers, S. Laurence, & Steven Stich (eds.): *The innate mind: Structure and contents*. New York: Oxford University Press, 107–121.

Samuels, Richard 2007. "Is innateness a confused notion?" in Peter Carruthers, S. Laurence & Steven Stich (eds.): *The Innate Mind: Foundations and the future*. New York: Oxford University Press.

Samuels, Richard 2009. "Nativism." In Sarah Robins, John Symons & Paco Calvo (eds.): *The Routledge Companion to Philosophy of Psychology*. Routledge.

Sansom, Roger & Robert N. Brandon (eds.) (2007): *Integrating Evolution and Development: From theory to practice*. Cambridge, Ma: The MIT Press.

Satz, Debra, & John Ferejohn (1994): "Rational choice and social theory", *Journal of Philosophy* 91: 71–87.

Schaffer, Jonathan (2005): "Contrastive causation", *The Philosophical Review* 114: 297–328.

Schaller, Mark, Ara Norenzayan, Steven J. Heine, Toshio Yamagishi & Tatsuya Kameda (eds.) (2010): *Evolution, Culture, and Human Mind*. New York: Psychology Press (Taylor & Francis Group).

Schelling, Thomas 1971. "Dynamic Models of Segregation", *Journal of Mathematical Sociology* 1: 143–186.

Schelling, Thomas 1978. *Micromotives and Macrobehavior*. New York: Norton.

Scher, Steven J. & Frederick Rausher (eds.) 2002. *Evolutionary Psychology: Alternative Approaches*. Kluwer, Dordrecht.

Schloss, Jeffrey & Michael Murray 2009. *The Believing Primate: Scientific, Philosophical, and Theological Reflections on the Origin of Religion*. Oxford: Oxford University Press.

Searcy, Christopher A., Levi N. Gray, Peter C. Trenham & H. Bradley Shaffer 2014. "Delayed life history effects, multilevel selection, and evolutionary trade-offs in the California tiger salamander", *Ecology* 95: 68–77.

Seemann, Axel (ed.) 2012. *Joint Attention: New Developments in Psychology, Philosophy of Mind, and Social Neuroscience*. MIT Press.

Segerstråle, Ullica 2000. *Defenders of Truth: The Sociobiology Debate*. Oxford: Oxford University Press.

Sehon, Scott 1997. "Deviant Causal Chains and the Irreducibility of Teleological Explanation", *Pacific Philosophical Quarterly* 78: 195–213.

Sehon, Scott 2005. *Teleological Realism: Mind, Agency, and Explanation*. Cambridge, MA: MIT Press.

Sellars, Wilfrid 1956. "Empiricism and the Philosophy of Mind". In Herbert Feigl & Michael Scriven (eds.): *Minnesota Studies in the Philosophy of Science, Volume I: The Foundations of Science and the Concepts of Psychology and Psychoanalysis.*University of Minnesota Press, Minnesota. 253–329.

Sesardic, Neven 1995. "Recent Work on Human Altruism and Evolution", *Ethics* 106: 128–157.
Sesardic, Neven 1999. "Altruism", *British Journal for the Philosophy of Science* 50: 457–466.
Shafer-Landau, R. 2003. *Moral Realism: A Defence*. Oxford: Clarendon Press.
Shamay-Tsoory, Simone G., Hagai Harari, Judith Aharon-Peretz & Yechiel Levkovitz 2010. "The role of the orbitofrontal cortex in affective theory of mind deficits in criminal offenders with psychopathic tendencies", *Cortex* 46, 668–677.
Shapiro, Lawrence & Elliot Sober 2007. "Epiphenomenalism – The Do's and Don'ts." In Gereon Wolters & Peter Machamer (eds.): *Thinking About Causes: From Greek Philosophy to Modern Physics*. Pittsburgh: University of Pittsburgh Press, 235–264.
Shea, Nicholas 2011. "Developmental Systems Theory Formulated as a Claim about Inherited Representations". *Philosophy of Science* 78: 60–82.
Shea, Nicholas 2013. "Two modes of transgenerational information transformation." In Kim Sterelny, Richard Joyce, Brett Calcott & Ben Fraser (eds.): *Cooperation and Its Evolution*. Cambridge, Ma: The MIT Press, 289–312.
Shea, Nicholas 2018. *Representation in cognitive science*. Oxford: Oxford University Press.
Sherman, Paul 1988. "The Levels of Analysis" *Animal Behavior* 36: 616–619.
Sherry, David 2005. "Do ideas about function help in the study of causation?" *Animal Biology* 55: 441–456.
Shettleworth, Sara J. 2010. *Cognition, Evolution, and Behavior*. Second edition. Oxford: Oxford University Press.
Silverman, Sydel 2002. "Foreword", in Richard G. Fox & Barbara J. King (eds.): *Anthropology Beyond Culture*. Oxford: Berg.
Sintonen, Matti 1984. *The Pragmatics of Scientific Explanation*. Acta Philosophica Fennica 37. Societas Philosophica Fennica. Helsinki.
Sintonen, Matti 1989. "Explanation: In Search of Rationale". In Wesley Salmon & Philip Kitcher (eds.): *Scientific Explanation*. Minnesota Studies in the Philosophy of Science, Vol. XIII. Minneapolis: University of Minnesota Press. 253–282.
Sintonen, Matti 1990. "How to Put Questions to Nature". In D. Knowles (ed.): *Explanation and its Limits*. Cambridge: Cambridge University Press. 267–284.
Skinner, B. F. 1957. *Verbal Behavior*. Copley Publishing Group.
Skipper, Robert & Roberta Milstein 2005. "Thinking about evolutionary mechanisms: natural selection", *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 327–347.

Sklar, Lawrence 2010. "I'd Love to Be a Naturalist – if Only I Knew What Naturalism Was", *Philosophy of Science* 77: 1121–1137.
Skyrms, Brian 1996. *Evolution of the Social Contract*. Cambridge: Cambridge University Press.
Smith, Eric Alden 2000. "Three styles in the evolutionary analysis of human behaviour", in Lee Cronk, Napoleon Chagnon & William Irons: *Adaptation and Human Behavior: An Anthropolgoical Perspective*. New York: Aldine De Gruyter, 27–48.
Smith, Eric Alden & Bruce Winterhalder (eds.) 1992. *Evolutionary Ecology and Human Behavior*. Hawthorne, NY: Aldine de Gruyter.
Smith, Eric Alden, Monique Borgerhoff Mulder & Kim Hill 2000. "Evolutionary analyses of human behaviour: a commentary on Daly & Wilson", *Animal Behaviour* 60: F21–F26.
Smith, Subrena E. 2020. "Is Evolutionary Psychology Possible?" *Biological Theory* 15: 39–49.
Sober, Elliot 1980a. "Evolution, population thinking, and essentialism", *Philosophy of Science* 47: 350–383.
Sober, Elliot 1980b. "Holism, Individualism, and the Units of Selection". In P. Asquith and R. Giere (eds.): *Proceedings of the Biennial Meetings of the Philosophy of Science Association* 2: 93–101.
Sober, Elliot 1984a. *The Nature of Selection: The Evolutionary Theory in Philosophical Focus*. The MIT Press, Cambridge, MA.
Sober, Elliot (ed.) 1984b. *Conceptual Issues in Evolutionary Biology*. The MIT Press, Cambridge, MA.
Sober, Elliot 1985. "Two Concepts of Cause", in Peter Asquith & Philip Kitcher (eds.), *PSA 1984*. East Lansing: Philosophy of Science Association, 405–424.
Sober, Elliot 1988a. "What is evolutionary altruism?" *Canadian Journal of Philosohy* 14: 75–99.
Sober, Elliot 1988b. "Apportioning causal responsibility", *Journal of Philosophy* 85: 303–318.
Sober, Elliot 1989. "Evolutionary altruism and psychological egoism", in J. E. Fenstad, I. T. Frolov, & R. Hilpinen (eds.): *Logic, Methodology and Philosophy of Science VIII*, Elsevier Science Publishers, Amsterdam, 495–514.
Sober, Elliot 1992a. "The evolution of altruism: correlation, cost, and benefit", *Biology and Philosophy* 7: 177–187.
Sober, Elliot 1992b. "Did evolution make us psychological egoists?" In Elliot Sober (1994): *From a Biological Point of View*, Cambridge University Press, Cambridge, 8–27.

Sober, Elliot 1992c. "Hedonism and Butler's Stone", *Ethics* 103: 97–103.
Sober, Elliot 1992d. "Screening-Off and the Units of Selection". *Philosophy of Science* 59: 142–152.
Sober, Elliot 1994a. *From a Biological Point of View: Essays in Evolutionary Philosophy*. Cambridge University Press, Cambridge.
Sober, Elliot 1994b. "Did Evolution Make Us Psychological Egoists?" In Elliot Sober: *From a Biological Point of View: Essays in Evolutionary Philosophy*. Cambridge University Press, Cambridge. 8–27.
Sober, Elliot 1994c. "Prospects for an Evolutionary Ethics". In Elliot Sober: *From a Biological Point of View: Essays in Evolutionary Philosophy*. Cambridge University Press, Cambridge. 93–113.
Sober, Elliot 1998a. "Innate Knowledge". In Craig (ed.): *Routledge Encyclopedia of Philosophy*. London: Routledge, 794–797.
Sober, Elliot 1998b. "Six Sayings about Adaptationism". In David L. Hull & Michael Ruse (eds.): *The Philosophy of Biology*. Oxford University Press, Oxford. 72–86.
Sober, Elliot 1998c. "Three differences between evolution and deliberation", in: Peter Danielson (ed.): *Modeling rationality, morality and evolution*. Oxford: Oxford University Press, 408–422.
Sober, Elliot 2001. "The two faces of fitness", in R. Singh, D. Paul, C. Krimbas, and J. Beatty (eds.): *Thinking about Evolution: Historical, Philosophical, and Political Perspectives*, Cambridge University Press, Cambridge, 309–321.
Sober, Elliot 2009. "Parsimony and models of animal minds", in Robert Lurz (ed.), *The Philosophy of Animal Minds*. New York: Cambridge University Press, 237–257.
Sober, Elliot & Richard C. Lewontin 1982. "Artifact, Cause and Genic Selection", *Philosophy of Science* 49: 157–180.
Sober, Elliot & David Sloan Wilson 1994. "A Critical Review of Philosophical Work on the Units of Selection Problem", *Philosophy of Science* 61: 534–55.
Sober, Elliot & David Sloan Wilson 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press, Cambridge, MA.
Sober, Elliot & David Sloan Wilson 2000a. "Summary of *Unto Others: The Evolution of Unselfish Behavior*", Journal of Consciousness Studies 7: 185–206.
Sober, Elliot & David Sloan Wilson 2000b. "Morality and *Unto Others*", *Journal of Consciousness Studies* 7: 257–268.
Sousa, Paulo, Scott Atran & Douglas Medin 2002. "Essentialism and folkbiology: Further evidence from Brazil", *Journal of Cognition and Culture* 2, 195–223.
Spears, Russell, Penelope Oakes, Naomi Ellemers & Alexander Haslam (eds.) 1997. *The Psychology of Stereotyping and Group Life*. London: Basil Blackwell.

Spencer, Herbert 1864. *Principles of Biology*. Volume I. London: Williams and Norgate.
Spencer, Herbert 1898 [1874–1896]. *Principles of Sociology*. New York: D. Appleton and Compnany.
Spencer, Herbert 1901 [1868–1874]. *Essays: Moral, Political and Speculative*. London: Williams and Norgate.
Sperber, Dan 1994. "The modularity of thought and the epidemiology of representations", in Lawrence A. Hirschfeld & Susan A. Gelman (eds.): *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge: Cambridge University Press.
Sperber, Dan 1996. *Explaining Culture: A Naturalistic Approach*. Blackwell.
Sperber, Dan 1997. "Individualisme méthodologique et cognitivisme." In R. Boudon, F. Chazel & A. Bouvier (eds.): *Cognition et sciences sociales*. Presse Universitaires de France, 123–136. The unpublished English version "Methodological individualism and cognitivism in the social sciences": https://www.dan.sperber.fr/?p=33
Sperber, Dan 2000. "An Objection to the Memetic Approach to Culture", in Robert Aunger (ed.): *Darwinizing Culture*. Oxford: Oxford University Press, 163–173.
Sperber, Dan 2001. "In Defense of massive modularity." In Emmanuel Dupoux (ed.): *Language, Brain and Cognitive Development: Essays in Honor of Jacques Mehler*. Cambridge, MA: MIT Press. 47–57.
Sperber, Dan, Francesco Cara & Vittorio Girotto (1995): "Relevance theory explains the selection task", *Cognition* 57, 31–95.
Sripada, Chandra Sekhar & Stephen Stich 2007. "A framework for the psychology of norms". In Peter Carruthers, Stephen Laurence & Stephen Stich (eds.): *The Innate Mind: Volume 2: Culture and Cognition*. Oxford: Oxford University Press.
Stanovich, Keith E. 1999. *Who is Rational? Studies of individual differences in reasoning*. Lawrence Erlbaum.
Stanovich, Keith E. 2011. *Rationality and the Reflective Mind*. Oxford: Oxford University Press.
Stanovich, Keith E. 2012. "On the Distinction between Rationality and Intelligence: Implications for Understanding Individual Differences in Reasoning", in Keith J. Holyoak & Robert G. Morrison (eds.): *The Oxford Handbook of Thinking and Reasoning*. Oxford: Oxford University Press, 343–365.
Stanovich, Keith E. & Richard F. West 2000. "Advancing the rationality debate", *Behavioral and brain sciences* 23: 701–717.

St Clair, James J. H. & Christian Rutz 2013. "New Caledonian crows attend to multiple functional properties of complex tools", *Philosophical Transactions of Royal Society of London B* 368: 20120415. doi: 10.1098/rstb.2012.0415.
Stephens, Christopher 2004. "Selection, Drift, and the "Forces" of Evolution", *Philosophy of Science* 71: 550–570.
Stephens, Christopher 2010. "Forces and causes in evolutionary theory", *Philosophy of Science* 77: 716–727.
Stephens, David & John Krebs 1986. *Foraging theory*. Princeton University Press, Princeton
Sterelny, Kim 1995. "Basic Minds", *Philosophical Perspectives* 9: 251–270.
Sterelny, Kim 1996a. "The return of the group", *Philosophy of Science* 63: 562-84.
Sterelny, Kim 1996b. "Explanatory Pluralism in Evolutionary Biology", *Biology and Philosophy* 11: 193–214.
Sterelny, Kim 1998. "Intentional Agency and the Metarepresentation Hypothesis", *Mind and Language* 13: 11–28.
Sterelny, Kim 1999. "Situated Agency and the Descent of Desire". In Valerie Gray Hardcastle (ed.): *Biology Meets Psychology: Constraints, Conjectures, Connections*. MITPress, Cambridge.
Sterelny, Kim 2001a. *The Evolution of Agency and Other Essays*. Cambridge University Press, Cambridge.
Sterelny, Kim 2001b. "Evolution and Agency". In Kim Sterelny: *The Evolution of Agency and Other Essays*. Cambridge University Press, Cambridge. 3–26.
Sterelny, Kim 2001c. "The Evolution of Agency". In Kim Sterelny: *The Evolution of Agency and Other Essays*. Cambridge University Press, Cambridge. 260–288.
Sterelny, Kim 2003. *Thought in a Hostile World: The Evolution of Human Cognition*, Blackwell Publishing, Oxford.
Sterelny, Kim 2006a. "Memes Revisited", *British Journal for the Philosophy of Science* 57: 145–165
Sterelny, Kim 2006b. "The Evolution and Evolvability of Culture", *Mind and Language* 21: 137–165.
Sterelny, Kim 2007. "An alternative evolutionary psychology?" in Gangestad & Simpson (eds.): *The Evolution of Mind*, New York: Guilford Press, 178–185.
Sterelny, Kim 2012. *The Evolved Apprentice: How Evolution Made Humans Unique*.
Sterelny, Kim 2013. "Living in interesting times: cooperation and collective action in the Holocene." In Kim Sterelny, Richard Joyce, Brett Calcott & Ben Fraser (eds.): *Cooperation and Its Evolution*. Cambridge, Ma: The MIT Press, 89–108.
Sterelny, Kim 2015. "Content, Control and Display: The Natural Origins of Content", *Philosophia* 43: 549–564.

Sterelny, Kim, Richard Joyce, Brett Calcott & Ben Fraser (eds.) 2013. *Cooperation and Its Evolution*. Cambridge, Ma: The MIT Press.
Sterelny, Kim & Philip Kitcher 1988. "The Return of the Gene", *Journal of Philosophy* 85: 339–361.
Sterelny, Kim, Kelly Smith & Mike Dickinson 1996. "The Extended Replicator", *Biology and Philosophy* 11: 377–403.
Stevens, Lori, Charles J. Goodnight & Susan Kalisz 1995. "Multilevel selection in natural populations of *Impatiens capensis*", *The American Naturalist* 145: 513–526.
Stich, Stephen 1975. "The Idea of Innateness", in Stich (ed.), *Innate Ideas*, Los Angeles: University of California Press.
Stich, Stephen 1978. "Beliefs and sub-doxastic states", *Philosophy of Science* 45: 499–518.
Stich, Stephen 1983. *From Folk Psychology to Cognitive Science: The Case Against Belief*. MIT Press.
Stich, Stephen 1996. *Deconstructing the Mind*. MIT Press.
Stich, Stephen 2007. "Evolution, altruism and cognitive architecture: A critique of Sober and Wilson's argument for psychological altruism", *Philosophy and Biology* 22: 267–281.
Stotz, Karola 2014. "Extended evolutionary psychology: the importance of transgenerational developmental plasticity". *Frontiers in Psychology* 5. doi:10.3389/fpsyg.2014.00908
Stotz, Karola & Paul Griffiths 2004: "Genes: Philosophical analyses put to the test", *History and Philosophy of the Life Sciences* 26: 5–28.
Stotz, Karola, Paul Griffiths & Rob Knight 2004. "How biologists conceptualize genes: An empirical study", *Studies in History and Philosophy of the Biological and Biomedical Sciences* 35: 647–673.
Strawson, Peter 1985. *Scepticism and Naturalism: Some Varieties*. Methuen.
Suchak, Malini, Timothy M Eppley, Matthew W. Campbell, Rebecca A. Feldman, Luke F. Quarles & Frans de Waal 2016. "How chimpanzees cooperate in a competitive world", *Proceedings of the National Academy of Sciences of the United States of America* 113: 10215–10220. doi:10.1073/pnas.1611826113.
Symons, Donald 1979. *The Evolution of Human Sexuality*. Oxford: Oxford University Press.
Symons, Donald 1989. "A critique of Darwinian anthropology", *Ethology and Sociobiology* 10: 131–143.
Tammela, T., G. Zarkada, E. Wallgard, A. Murtomäki, S. Suchting, M. Wirzenius, M. Waltari, M. Hellström, T. Schomber, R. Peltonen, C. Freitas, A. Duarte, H. Isoniemi, P. Laakkonen, G. Christofori, S. Ylä-Herttuala, M.

Shibuya, B. Pytowski, A. Eichmann, C. Betsholtz & K. Alitalo 2008. "Blocking VEGFR-3 suppresses angiogenic sprouting and vascular network formation", *Science* 454: 656-660.
Taylor, Peter D. & Andrew J. Irwin 2000. "Overlapping generations can promote altruistic behavior", *Evolution* 54: 1135–1141.
Taylor, Ridchard 1966. *Action and Purpose*. Englewood Cliffs: Prentice-Hall.
Thalos, Miriam 2013. *Without hierarchy: The scale freedom of the universe*. Oxford: Oxford University Press.
Thoday, J. M., & A. S. Parkes (eds.) 1968. *Genetic and Environmental Influences on Behaviour*. New York: Plenum.
Tiger, Lionel, & Robin Fox 1971. *The Imperial Animal*. New York: Holt, Rinehart and Winston.
Tikhodeyev, Oleg 2018. "The mechanisms of epigenetic inheritance: How diverse are they?" *Biological Reviews of the Cambridge Philosophical Society* 93: 1987–2005.
Tinbergen, Nikolaas 1963. "On the aims and methods of ethology", *Zietschrift für Tierpsychologie* 20, 410–433.
Tomasello, Michael 1999. *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.
Tomasello, Michael 2008. *Origins of human communication*. Cambridge, MA: MIT Press.
Tomasello, Michael 2009. *Why We Cooperate*. Cambridge, Mass.: MIT Press.
Tomasello, Michael 2014. *A Natural History of Human Thinking*. Cambridge, MA: Harvard University Press.
Tomasello, Michael, Josep Call & Brian Hare 2003. "Chimpanzees understand psychological states – The question is which ones and to what extent", *Trends in Cognitive Sciences* 7: 153–156.
Tomasello, Michael, Malinda Carpenter, Josep Call, Tanya Behne & Henrike Moll 2005. "Understanding and sharing intentions: the ontogeny and phylogeny of cultural cognition", *Behavioral and Brain Sciences* 28: 675–735.
Tomkins, Silvan 1987. "Script Theory", in Joel Arnoff, A. I. Rabin, and Robert A. Zucker (eds.): *The Emergence of Personality*. Springer, New York, 147–216.
Tooby, John & Leda Cosmides 1989. "Evoutionary Psychology and the Generation of Culture, part I: Theoretical Considerations", *Ethology and Sociobiology* 11: 113–129.
Tooby, John & Leda Cosmides 1990a. "The past explains the present: emotional adaptations and the structure of ancestral environments", *Ethology and Sociobiology* 11: 375–424.

Tooby, John & Leda Cosmides 1990b. "On the universality of human nature and the uniqueness of the individual: the role of genetics and adaptation", *Journal of Personality* 58: 17–67.
Tooby, John & Leda Cosmides 1992. "The psychological foundations of culture." In Jerome Barkow, Leda Cosmides & John Tooby (eds.): *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, New York.
Tooby, John & Irven DeVore 1987. "The reconstruction of hominid behavioral evolution through strategic modelling", in Warren G. Kinzey (ed.): *The evolution of human behavior: Primate models*, Albany, NY: SUNY Press, 183–237.
Tripodi, Vera 2011. "What kind of intuition? Men, women, and philosophical reasoning", *Epistemologia* 34: 187–209.
Trivers, Robert L. 1971. "The evolution of reciprocal altruism", *Quarterly Review of Biology* 46: 35–57.
Trivers, Robert L. 1974. "Parent-offspring conflict", *American Zoologist* 14: 249–264.
Trivers, Robert L. 1985. *Social Evolution*. Menlo Park: Benjamin Cumins.
Tuomela, Raimo 1980. "Explaining Explaining", *Erkenntnis* 15: 211–243.
Tuomela, Raimo 1995. *The Importance of Us: A Study of Basic Social Notions.* Stanford: Stanford University Press.
Tuomela, Raimo 2007. *The Philosophy of Sociality: The Shared Point of View*. New York: Oxford University Press.
Tuomela, Raimo & Kaarlo Miller 1988. "We-Intentions", *Philosophical Studies* 53: 367–389.
Uhlmann, Eric Luis, David A. Pizarro & Paul Bloom 2008. "Varieties of Social Cognition", *Journal for the Theory of Social Behaviour* 38: 293–322.
Uller, Tobias, Armin P. Moczek, Richard A. Watson, Paul M. Brakefield & Kevin N. Laland 2018. "Developmental Bias and Evolution: A Regulatory Network Perspective". *Genetics* 209: 949–966.
van Baalen, Minus & David A. Rand 1998. "The Unit of Selection in Viscous Populations and the Evolution of Altruism", *Journal of Theoretical Biology* 193: 631–648.
van Fraassen, Bas 1980. *Scientific Image*. Oxford University Press. Oxford.
Vihvelin, Kadhri 1995. "Causes, Effects, and Counterfactual Dependece", *Australasian Journal of Philosophy* 73: 560–583.
von Wright, Georg Henrik 1971. *Explanation and Understanding*. Cornell University Press.
von Wright, Georg Henrik 2001. *In the Shadow of Descartes*. Springer.
Voorzanger, Bart 1994. "Bioaltruism reconsidered", *Biology and Philosophy* 9, 75–84.

Vosniadou, Stella, & Andrew Ortony (eds.) 1989. *Similarity and Analogical Reasoning*. Cambridge: Cambridge UP.
Vygotsky, Lev 1978. *Mind in society: the development of higher psychological processes*. Cambridge, MA: Harvard University Press.
Waddington, Conrad H. 1942. "Canalization of development and the inheritance of acquired characters". *Nature* 150: 563–565.
Waddington, Conrad H. 1952. "The Evolution of Developmental Systems". In D.A. Herbert (ed.): *Twenty-eighth Meeting of the Australian and New Zealand Association for the Advancement of Science, Brisbane, Australia*. Brisbane: A.H Tucker, Government Printer.
Waddington, Conrad H. 1953. "Genetic assimilation of an acquired character". *Evolution* 7: 118–126.
Waddington, Conrad H. 1957. *The Strategy of the Genes*. George Allen & Unwin.
Wagner, Andreas 2005. *Robustness and Evolvability in Living Systems*. Princeton Studies in Complexity. Princeton: Princeton University Press.
Wagner, Wolfgang & Günter Wagner 2003. "Examining the modularity concept in evolutionary psychology: The level of genes, mind and culture", *Journal of Cultural and Evolutionary Psychology* 1.
Wakefield, Jerome C. 1992. "The concept of mental disorder: On the boundary between biological facts and social values", *American Psychologist* 47: 373–388.
Wallace, Bruce 1968. "Polymorphism, population size, and genetic load". In Richard C. Lewontin (ed.): *Population Biology and Evolution*. Syracuse University Press. 87–108.
Walsh, Dennis M. 1996. "Fitness and Function", *British Journal for the Philosophy of Science* 47: 553–574.
Walsh, Denis M. 2000. "Chasing Shadows – Natural Selection and Adaptation", *Studies in the History and Philosophy of Biology and the Biomedical Sciences* 31: 135–153.
Walsh, Denis M. & Andre Ariew 1996. "A taxonomy of functions", *Canadian Journal of Philosophy* 26: 493–514.
Walsh, Denis M., Andre Ariew & Tim Lewens 2002. "The trials of life: Natural selection and random drift", *Philosophy of Science* 69: 452–473.
Wang, Qi 2017. "Why Should We All Be Cultural Psychologists? Lessons from the Study of Social Cognition". *Perspectives in Psychological Science* 11: 583–596.
Warneken, Felix 2013. "Altruistic behaviors from a developmental and comparative perspective." In Kim Sterelny, Richard Joyce, Brett Calcott & Ben Fraser (eds.): *Cooperation and Its Evolution*. Cambridge, Ma: The MIT Press, 399–423.

Wason, Peter C. 1966. "Reasoning", in Foss, B. M. (ed.): *New horizons in psychology*. Harmondsworth: Penguin.
Weinberg, Jonathan, & Stephen Crowley 2009. "The x-phi(les): unusual insights into the nature of inquiry", *Studies in History and Philosophy of Science* 40: 227–232.
Weinberg, Jonathan & Ron Mallon 2008. "Living with innateness (and environmental dependence too)", *Philosophical Psychology* 21: 415–424.
Weinberg, Jonathan, Shaun Nichols & Stephen Stich 2001. "Normativity and Epistemic Intuitions", *Philosophical Topics* 29: 429–460.
Weinig, Cynthia, Jill A. Johnston, Charles G. Willis & Julin N. Maloof 2007. "Antagonistic multilevel selection on size and architecture in variable density settings", *Evolution* 61: 58–6.
Wellman, Henry M. 1990. *The Child's Theory of Mind*. Cambridge, MA: MIT Press.
Wellman, Henry M., Donna Cross & Julie Watson 2001. "Meta-analysis of Theory-of-mind Development: The Truth about False Belief", *Child Development* 72: 655–684.
Wellman, Henry M. & David Liu 2004. "Scaling of Theory-of-Mind Tasks", *Child Developing* 75: 523–541.
West, Stuart A. & Andy Gardner 2010. "Altruism, spite and greenbeards." *Science* 327: 1341–1344.
West, Stuart A. & Andy Gardner 2013. "Adaptation and inclusive fitness." *Current Biology* 23: R577–584.
West, Stuart A., Ashleigh S. Griffin & Andy Gardner 2007a. "Evolutionary explanations for cooperation." *Current Biology* 17: R661–672.
West, Stuart A., Ashleigh S. Griffin & Andy Gardner 2007b. "Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection." *Journal of Evolutionary Biology* 21: 374–385.
West, Stuart A., Claire El Mouden & Andy Gardner 2001. "Sixteen common misconceptions about the evolution of cooperation in humans." *Evolution and Human Behavior* 32: 231–262.
West-Eberhard, Mary Jane 1992. "Adaptation: Current Usages". In Evelyn Fox Keller & Elizabeth Lloyd (eds.): *Keywords in Evolutionary Biology*. Harvard University Press, Cambridge, MA. 13–18.
West-Eberhard, Mary Jane 2003. *Developmental Plasticity and Evolution*. Oxford: Oxford University Press.
West-Eberhard, Mary Jane 2005. "Phenotypic accommodation: adaptive innovation due to developmental plasticity". *Journal of Zoology Part B: Molecular and Developmental Evolution* 304: 610–618.

Whiten, Andrew 1996. "When does smart behaviour-reading become mind-reading?" In Peter Carruthers and Peter K. Smith (eds.) *Theories of Theories of Mind*. Cambridge: Cambridge University Press. 277–292.

Whiten, Andrew 2009. "The identification and differentiation of culture in chimpanzees and other animals: from natural history to diffusion experiments", in Kevin N. Laland & Bennett G. Galef (eds.): *The question of animal culture*. Cambridge: Harvard University Press, 99–124.

Whiten, Andrew, & Richard W. Byrne (eds.) 1997. *Machiavellian Intelligence, Vol. 2: Evaluations and Extensions*. Cambridge: Cambridge University Press.

Whiten, Andrew, Victoria Horner & Sarah Marshall-Pescini 2003. "Cultural panthropology", *Evolutionary Anthropology: Issues, News, and Reviews* 12: 92–105.

Whiten, Andrew, Nicola McGuigan, Sarah Marshall-Pescini & Lydia M. Hopper 2009. "Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee",

Wilkes, Kathleen 1988. *Real People: Personal Identity without Thought Experiments*. Oxford: Oxford University Press.

Wilkins, John & Paul Griffith 2012. "Evolutionary debunking arguments in three domains: Fact, value, and religion", In James Maclaurin Greg Dawes (ed.), A New Science of Religion. Routledge.

Williams, George C. 1966. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princetown University Press.

Wilson, David Sloan 1975. "A theory of group selection", *Proceedings of the National Academy of Sciences* 72: 143–146.

Wilson, David Sloan 1989. "Levels of Selection: An Alternative to Individualism in Biology and the Human Sciences". *Social Networks* 11: 257–272.

Wilson, David Sloan 1992. "On the relationship between evolutionary and psychological definitions of altruism and selfishness", *Biology and Philosophy* 7: 61–68.

Wilson, David Sloan 1997. "Incorporating group selection into the adaptationist programme: A case study involving human decision making". In Jeffry A. Simpson & Douglas Kenrick (eds.): *Evolutionary Social Psychology*. Hillsdale: Erlbaum. 348–386.

Wilson, David Sloan 2002. *Darwin's Cathedral: Evolution, Religion, and the Nature of Society*, The University of Chicago Press, Chicago.

Wilson, David Sloan & Elliot Sober 1989. "Reviving the superorganism", *Journal of Theoretical Biology* 136: 337–356.

Wilson, David Sloan & Elliot Sober 1994. "Reintroducing group selection to the human behavioral sciences", *Behavioral and Brain Sciences* 17(4): 585–608.

Wilson, David Sloan, Carolyn Wilczynski, Alexandra Wells & Laura Weiser 2000. "Gossip and other aspects of language as group-level adaptions". In Celia Heyes & Ludwig Huber: *Evolutionary Social Psychology*. Cambridge: MIT Press. 347–365.

Wilson, Edward O. 1975. *Sociobiology: The New Synthesis*, Harvard University Press, Cambridge, MA.

Wilson, Edward O. 1978. *On Human Nature*. Harvard University Press, Cambridge, MA.

Wilson, Jack 2002. "Accidental altruist: biological analogues for intention", *Biology and Philosophy* 17: 71–91.

Wilson, Robert (ed.) 1999. *Species. New interdisciplinary essays*. Cambridge, MA: MIT Press.

Wimsatt, William C. 1974. "Complexity and organization". In K. F. Schaffner & R. S. Cohen (eds.); *PSA 1972*. Dordrecht, NL: Reidel. 67–86.

Wimsatt, William C. 1976. "Reductionism, levels of organization, and the mind-body problem". In G. Globus, I. Savodnik & G. Maxwelll (eds.): *Consciousness and the brain*. New York: Plenum Press. 199–267.

Wimsatt, William C. 1980. "Reductionistic research strategies and their biases in the units of selection controversy". In Thomas Nickles (ed.): *Scientific Discovery: Case Studies.* Boston Studies in the Philosophy of Science 60: 213–259. Dordrecht: Springer.

Wimsatt, William C. 1986a. "Developmental Constraints, Generative Entrenchment and the Innate-Aquired Distinction". In W. Bechtel (ed.): *Integrating Scientific Disciplines*. Dordrecht: Martinus Nijhoff, 185–208.

Wimsatt, William C. 1986b. "Forms of Aggregativity". In Alan Donagan, Anthony N. Perovich Jr. & Michael V. Wedin (eds.): *Human Nature and Natural Knowledge. Essays Presented to Marjorie Grene on the Occasion of Her Seventy-Fifth Birthday*. Boston Studies in the Philosophy of Science 89. Dordrecht: Springer. 259–291.

Wimsatt, William C. 1997. "Aggregativity: Reductive heuristics for finding emergence". *Philosophy of Science* 64: 372–384.

Wimsatt, William C. 1999. "Generativity, Entrenchment, Evolution, and Innateness: Philosophy, Evolutionary Biology, and Conceptual Foundations of Science". In V. G. Hardcastle (ed.): *Where Biology Meets Psychology: Philosophical Essays*. Cambridge, Mass.: MIT Press, 139–179.

Wimsatt, William C. 2001. "Generative entrenchment and the developmental systems approach to evolutionary process". In Susan Oyama, Paul E.

Griffiths & Russell D. Gray (eds.): *Cycles of Contingency: Developmental Systems and Evolution*. Cambridge, Ma: MIT Press. 219–238.
Wimsatt, William C. 2007. *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Harvard University Press.
Wimsatt, William C. & Jeffrey C. Schank 1988. "Two constraints on the evolution of complex adaptations and the means of their avoidance". In Matthew H. Nitecki (ed.): *Evolutionary Progress*. Chicago: University of Chicago Press, 231–273.
Winkielman, Pitor & Kent Berridge 2004. "Unconscious Emotion", *Current Directions in Psychological Science* 13: 120–123.
Wittgenstein, Ludwig 1953. *Philosophical Investigations*. (Translated from a German manuscript by Elizabeth Anscombe.) Macmillan Publishing Company.
Wolf, Jason B., Edmund D. Brodie III & Allen J. Moore 1999. "Interacting Phenotypes and the Evolutionary Process. II. Selection Resulting from Social Interactions", *American Naturalist* 153: 254–266.
Wolpert, Lewis, Cheryl Tickle, Peter Lawrence, Elliot Meyerowitz, Elizabeth Robertson, Jim Smith & Thomas Jessell 2010. *Principles of Development*, Fourth Edition. Oxford: Oxford University Press.
Wolters, Gereon & James G. Lennox (eds.) 1995. *Concepts, Theories, and Rationality in the Biological Sciences*. Pittsburgh: University of Pittsburgh Press.
Woodward, James 2000. "Explanation and invariance in the special sciences", *British Journal for the Philosophy of Science* 51: 197–254.
Woodward, James 2001. "Law and explanation in biology: invariance is the kind of stability that matters", *Philosophy of Science* 68: 1–20.
Woodward, James 2002. "What is a mechanism? A counterfactual account", *Philosophy of Science* 69 (Proceedings): S366–377.
Woodward, James 2003a. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.
Woodward, James 2003b. "Experimentation, causal inference, and instrumental realism." In Hans Radder (ed.): *The Philosophy of Scientific Experimentation*. Pittsburgh: University of Pittsburgh Press, 87–118.
Woodward, James 2004. "Counterfactuals and causal explanation", *International Studies in Philosophy of Science* 18: 41–72.
Woodward, James 2008. "Mental Causation and Neural Mechanisms". In Jakob Hohwy & Jesper Kallestrup (eds.): B*eing Reduced: New Essays on Reduction, Explanation, and Causation*. Oxford: Oxford University Press.
Woodward, James 2010. "Causation in biology: stability, specificity, and the choice of levels of explanation", *Biology and Philosophy* 25: 287–318.

Woodward, James 2011. "Mechanisms revisited", *Synthese* 183: 409–427.
Woodward, James 2014. "Interventionism and Causal Exclusion", *Philosophy and Phenomenological Research* (Early View, article first published online: 7 APR 2014, doi: 10.1111/phpr.12095)
Woodward, James, & Fiona Cowie 2004. "Is the mind a system of modules shaped by natural selection?" In Christoper Hitchcock (ed.) 2004. *Contemporary Debates in Philosophy of Science*. Oxford: Blackwell. 312–334.
Woodward, James, & Christoper Hitchcock 2003. "Explanatory generalizations, part I: A counterfactual account", *Nôus* 37: 1–24.
Wouters, Arno 1995. "Viability explanation", *Biology and Philosophy* 10: 435–457.
Wouters, Arno 2003. "Four notions of biological function", *Studies in the History and Philosophy of Biological and Biomedical Sciences* 34: 633–668.
Wouters, Arno 2005. "The functional perspective in organismic biology". In T. A. C. Reydon & L. Hemerik (eds.): *Current themes in theoretical biology*. Dordrecht: Springer. 33–69.
Wouters, Arno 2013. "Biology's functional perspective: Roles, advantage, and organization." In K. Kampourakis (ed.): *The philosophy of biology: A companion for educators*. Dordrecht: Springer. 455–486.
Wright, Larry 1973. "Functions", *Philosophical Review* 82: 139–168.
Wright, Larry 1976. *Teleological Explanations*. Berkeley: University of California Press.
Wynne-Edwards, Vero Copner 1962. *Animal Dispersion in Relation to Social Behaviour*, Oliver and Boyd, Edinburgh.
Wynne-Edwards, Vero Copner 1964. "Group selection and kin selection: response to Maynard-Smith", *Nature* 201: 1145–1147.
Ylikoski, Petri 2001. *Understanding Interests and Causal Explanation*, Academic Dissertation, University of Helsinki. http://urn.fi/URN:ISBN:951-45-9942-X
Ylikoski, Petri 2007. "The Idea of Contrastive Explanandum", in Johannes Persson & Petri Ylikoski (eds.): *Rethinking Explanation*. Dordrecht: Springer, 27–42.
Ylikoski, Petri 2009. "The illusion of depth of understanding in science". In H. de Regt, S. Leonelli & K. Eigner (eds.): *Scientific Understanding: Philosophical Perspectives*. Pittsburgh: University of Pittsburgh Press.
Ylikoski, Petri 2012. "Micro, macro, and mechanisms". In Harold Kincaid (ed.), *The Oxford handbook of philosophy of the social sciences*. Oxford: Oxford University Press. 21–45.
Ylikoski, Petri 2017. "Methodological Individualism", in L. McIntyre & A. Rosenberg (eds.): *Routledge Companion to Philosophy of Social Science*. New York: Routledge, 135–146.

Ylikoski, Petri & Emrah Aydinonat 2014. "Understanding with Theoretical Models". *Journal of Economic Methodology* 21: 19–36.
Ylikoski, Petri & Tomi Kokkonen 2009. *Evoluutio ja ihmisluonto*. Helsinki: Gaudeamus.
Ylikoski, Petri & Jaakko Kuorikoski 2010. "Dissecting explanatory power", *Philosophical Studies* 148: 201–219.
Ylikoski, Petri & Jaakko Kuorikoski 2016. "Self-interest, norms, and explanation", in Mark Risjord (ed.): *Normativity and Naturalism in the Philosophy of the Social Sciences*. New York: Routledge, 212–229.
Yzerbyt, Vincent, Olivier Corneille & Claudia Estrada 2001. "The interplay of subjective essentialism and entitativity in the formation of stereotypes", *Personality and Social Psychology Review* 5, 141–155.
Zahavi, Dan 2008. "Simulation, projection and empathy", *Consciousness and Cognition* 17: 514–522.
Zahle, Julie & Finn Collin (eds.) 2014a. *Rethinking the Individualism-Holism Debate: Essays in the philosophy of social science*. Dordrecht: Springer.
Zahle, Julie & Finn Collin 2014b. "Introduction", in Julie Zahle & Finn Collin (eds.): *Rethinking the Individualism-Holism Debate: Essays in the philosophy of social science*. Dordrecht: Springer, 1–14.
Zawidzki, Tad 2013. *Mindshaping: A New Framework for Understanding Human Social Cognition*. Cambridge, Ma: MIT Press.
Zittoun, Taniaand & Alex Gillespie 2015. "Internalization: how culture becomes mind". *Culture & Psychology* 21: 477–491.