The interpretation of artifacts : a critique of Dennett's design stance

DOI:
10.6100/IR636331

Document status and date:
Published: 01/01/2008

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

The Interpretation of Artifacts

A Critique of Dennett's Design Stance

Melissa van Amerongen

Simon Stevin Series in the Philosophy of Technology

The Interpretation of Artifacts
A Critique of Dennett's Design Stance

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr.ir. C.J. van Duijn, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op dinsdag 19 augustus 2008 om 16.00 uur

door

Melissa van Amerongen

geboren te Amsterdam

Dit proefschrift is goedgekeurd door de promotor:

prof.dr.ir. A.W.M. Meijers

Copromotor:
prof.dr. C.F.R. Illies

Samenstelling promotiecommissie

Rector Magnificus, voorzitter
prof.dr.ir. A.W.M. Meijers, Technische Universiteit Eindhoven, Promotor
prof.dr. C.F.R. Illies, Technische Universiteit Eindhoven, Co-promotor
prof.dr. H. Kornblith, University of Massachusetts (US)
prof.dr.ir. P.A. Kroes, Technische Universiteit Delft
prof.dr. P.M.F. Oomen, Technische Universiteit Eindhoven
prof.dr. M.V.P. Slors, Radboud Universiteit Nijmegen

Research for this thesis was made possible by the Netherlands Organization for Scientific Research (NWO).

Van Amerongen, M.
The Interpretation of Artifacts; A Critique of Dennett's Design Stance
Den Texstraat 42 II, 1017 ZC, Amsterdam
E: m.v.amerongen@tue.nl
ISBN: 978-90-386-1339-0
ISSN: 1574-941X

Contents

Acknowledgements

So this is it. About five years of full time work, more than 200 pages, many cancellations of appointments with family and friends, even more terrible moods, and no prospect of being able to contribute to building a spaceship with it or curing a terrible disease. The result: the discovery of a contradiction. From that perspective, completing a philosophical book is not so much a test of competence, but rather a test of perseverance.

There are many people that I want to thank for their support and help. My colleagues at the Philosophical Department. Anthonie Meijers and Christian Illies, for supervision and guidance, for interesting philosophical discussions in unusual places, and for helping me get the dissertation to the finish. I am certainly not the easiest person to supervise. Rianne Schaaf for helping me get the manuscript to the publisher in time. Palmyre Oomen for her sharp and detailed criticism on the final manuscript and the long and deeply motivating discussions that followed. Giacomo and Krist for their tough criticism, essential input, moral support and most of all, for being such great company. I feel terribly guilty that I beat both of you to it. Beers are on me next time.

Staff and students of the department of Philosophy at Amherst, for a great time that I will never forget. Lynne Baker and Hilary Kornblith for their hospitality, and excellent feedback. Seamus, with whom I unfortunately lost contact.

Emile, Sabine, Sofie, Conor, Arnold, Rogier, Sam, Jeroen, and my sister Floor. You're not only great friends, but really helped me during some difficult times when I couldn't see the end coming. Your repertoire was as diverse as it was effective. Last minute editorial advise, teasing, compassionate looks of recognition, newborn seals, cynical, relativizing statements about science in general and philosophy in particular, chocolate-with-orange with beer and Idols, excellent dinners, kinky pink suits, and calling my cards when you should not have.

My parents Liesbeth and Ben for putting up with me and their patience. Juliette and Norfried for extra parental care, Ineke and Bert for always being close by. The Kennisnet research department, Alfons, Ronald, Hilda, Nadine and Jackie, for time and a free mind to finish the dissertation.

And Take, for still being my boy friend. You're crazy. Let's get old together.

1 Introduction

Technology is an important phenomenon in human life, and in human action. Technical artifacts mediate and influence our actions, enlarge our capacities, and they are almost omnipresent in our environment. We deal and act with them all the time. We recognize things as coffeepots, pencils, computers and traffic signs, usually without any effort. Technology thus is a pervasive part of our everyday world.

Philosophers agree, I suppose, that technology plays such a large role in human life. But it is not so clear what a philosopher can contribute to our understanding of technology. The continental tradition, interested in cultural and societal developments, is relatively well equipped to reflect upon technological developments in the broad sense (Heidegger: 1977, Jonas: 1979, Latour: 1992, Feenberg: 1999). The analytic or Anglo-Saxon tradition, however, is only hesitantly starting to pick up some interest in technology. Even though I am sure that there are some interesting connections to be made between the continental and analytic tradition, this thesis focuses exclusively on the latter; acknowledging that the borders between the two traditions are not only always sharp[1]. So in the rest of the thesis, whenever I am discussing the field of Philosophy of Technology, I will be referring to the Anglo-Saxon tradition.

Philosophers working in the Anglo-Saxon tradition traditionally divide the world into the realm of objects and the realms of agents, or more generally, into the *space of causes* and *the space of reasons* (see, e.g. McDowell: 1994). The space of designed or technical objects, however, has been largely neglected, while it is not at all clear whether the space of designed objects fits nicely into either one of the two categories. On the one hand, technical artifacts have a structure, and are constructed in a certain way such that they are capable of doing what they do: The space of causes. On the other hand, it is often argued, physics can not wholly explain technical artifacts. From the perspective of physics, a hammer is merely a piece consisting of wood and so on, but not a *hammer*. What does make

[1] A good example is the influence of phenomenology on the Anglo-Saxon tradition. Philosophers like John Searle, Don Ihde, Richard Rorty, and Jürgen Habermas, to name the obvious examples, are hard to place in either one of the traditions.

it a *hammer*? Philosophers disagree, but the answer seems to be somehow connected with philosophically heavily debated issues such as intentionality, purposiveness and normativity. It may be a *hammer* because it is conceived as such by us, because of our valuing them as a hammer, or perhaps because it fulfills the purpose of a hammer well. Technical artifacts, being *designed* objects, are created and used for a purpose by agents. This would place them rather in the space of reasons.

These kinds of issues have been picked up by the TU Delft program *The Dual Nature of Technical Artifacts*, a program that has inspired the research program at the TU Eindhoven (and, thus, this dissertation). The starting point of the Dual Nature project is the claim that technical artifacts are an interesting subject in their own right, because they are physical objects created with a *purpose* and are thus only understood properly if we take into account both their physical and their intentional nature. 'Function' is seen as the "bridging concept that relates the physical and intentional domain" and the aim is "to conceptualize the dual nature of technical artifacts via the concept of function" (Kroes and Meijers: 2006, 2). The claim is that technical artifacts have to be conceptualized *both* in terms of physics *and* of intentionality, in order to be complete.

The Dual Nature account is only the start of a possible solution. It provides a direction, but it does not yet *explain* how the notion of function is supposedly able to bridge mind and world. And this problem is of course closely related to the age-old problem in the philosophy of mind how intentionality fits the physical world, a.k.a. the mind-body problem. The debate over what technical artifacts, and their functions, are, in that sense plays out directly in the philosophy of mind. Those who believe that artifacts have to be understood at least partly as somehow intentional phenomena, will at some point have to deal with the question what intentions are and how they relate to the physical world. As such, the Dual Nature approach reemphasizes a big philosophical problem by pointing at the dual nature of technical artifacts, but does not solve it.

If we accept the claim of the Dual Nature project that technical artifacts do have an intentional nature, we should first understand how they could have this intentional nature at all. A possible direction is sometimes given in terms of 'original' and 'derived' intentionality (Searle: 1992, Haugeland: 1998). Original intentionality explains why things that do not have minds, can still have an intentional nature: they have intentionality *derivitably*. A common example of derived intentionality is a written sentence. A sentence means something, but it *only* has meaning because *we*, bearers of original intentionality, attach meaning to it. Without the original intentionality of human beings, the claim is, there would only be dots on a paper, or pixels on a computer screen, but not a meaningful sentence. The meaning of a text thus derives from us. Another example is money: the euro coin in my wallet only has value in relation to us, intentional human beings (cf. Searle: 1996).

The distinction between original and derived intentionality can also be used for understanding technical artifacts: functions of artifacts are derived from the 'original intentions' of designers and users. So if we want to know what the function of an artifact is - what its 'meaning' (in the sense of purpose) is - we have to appeal to the intentionality of the designer of the artifact, or its users.

Many philosophers agree that technical artifacts do not fit easy in the space of causes, or physics. But this common intuition can lead to quite different conclusions. One might conclude that artifacts need (also) to be embedded in a theory of intentionality in order to be understood. *Or* one might conclude that artifacts are no proper objects of investigation, exactly because they are phenomena that have a mind-dependent and thus (inter)subjective element. The problem of mind-dependence can be explained by means of the very simple thought that if there were no minds, a mind-dependent item would not exist. Mind-dependence thus makes a thing ontologically (qua existence) awfully suspicious. Furthermore, explaining artifacts in terms of some mysterious property of human minds – original intentionality – is for many philosophers not a satisfying solution.

The apparent mind-dependent nature of technical artifacts, I believe, is an important reason that mainstream philosophy has largely ignored the philosophy of technology and theories of artifact function. Naturalistic-minded philosophers of mind, that dominate the mainstream, regard technical artifacts, being so much related to human intentionality, to be non-natural kinds. They have either settled for a non-intentional account of function, or have discarded the topic for being unscientific and vague. Or they assume that once the hard problems of human intentionality are solved, artifacts will just follow trivially[2].

[2] Recently, there have been a number of publications that indicate that a critical exchange may be forthcoming and that, slowly, philosophers are picking up the topic of artifacts (Margolis and Laurence: 2007, Thomasson: 2007, Baker: 2007). I expect the number of publications on artifacts to rise in the coming years, especially in ontology and semantics.

An important voice in this still relatively isolated debate is Daniel Dennett. I consider Dennett to be one of the greatest and interesting philosophers of the 20th century. Dennett was taught philosophy in the analytic tradition (by Gilbert Ryle and Willard von Orman Quine), and has developed their ideas considerably into a grand and thought provoking naturalistic theory of mind, consciousness, and evolution, shaking up age old philosophical ideas, and establishing links between philosophy and the sciences. His theory has gained a lot of attention in the philosophy of mind and consciousness, theories of evolution, the philosophy of science and in debates about functions.

Dennett is one of the few naturalists who does have a theory of functions that can capture more than just their physical nature. He believes that artifacts are, certainly, dependent on our minds, but he does not have to appeal to original intentionality.

The details of this position will be laid out in detail in this thesis, but let me shortly sketch the contours of this idea. In order to do that, we need to understand a little more about Dennett's theory of mind. A long-standing question in the philosophy of mind is: What are minds and do they actually exist? Dennett's radical and provocative answer to this question is that minds do exist, but only from a certain external perspective, an interpreting stance which he calls the intentional stance. His theory of the intentional stance tries to unite two traditions that are usually considered to be incompatible: behaviorism and hermeneutics. From behaviorism, Dennett uses the idea that we can only make meaningful statements about observable behavior. But according to Dennett, we can *interpret* this behavior in intentional terms.

Dennett contrasts the intentional stance with the physical stance and the design stance. A stance can be seen as a certain way or strategy of looking at an object or system, that helps predict their behavior. From the physical stance we explain and predict the behavior of *physical objects*, in physical terms and on the basis of physical laws, like physicists do. When we take the design stance, we explain the workings of a *designed object* using our knowledge of the design of the system. For instance, as a user of an alarm clock, I do not need to know anything about the internal workings of the clock, to have it wake me. The design stance is thus in some cases more efficient than the physical stance. Obviously, this stance only works for designed objects, and only if they work as they should. Thirdly, there is the intentional stance. When we take the intentional stance, we try to explain and predict the behavior of a system as if it were a *rational agent*, by seeing it as being guided by a mind. This is an even more efficient stance, even though it is clearly not as exact as the physical stance and even the design stance.

Let me note that the intentional stance does not only work for human beings, but also for all kinds of other systems. A rather uncontroversial example from Dennett is that people, when playing check against a computer, tend to treat it as an intentional system, for instance by ascribing certain goals (winning, not losing the Queen) to it. But Dennett sees a much wider range of systems that we can effectively see from the intentional stance. Even very simple creatures, like the mot that finds itself its way by following the light, is interpretable from the intentional stance, as is the thermostat that 'believes it to be too cold in the room, and therefore orders the boiler to heat up the water'. For Dennett these are all intentional systems, because they are all interpretable in terms of intentional states.

What makes Dennett especially interesting for my purposes are his ideas about the design stance. The design stance is, as just explained, a way of looking at design (design is in that sense mind dependent), but it also fulfills a role of *explaining* minds.

Like the intentional stance, the design stance has a very broad range. Not only technical artifacts, but also biological items, such as hearts and legs, as products of a process that resembles a design process (i.e. the process of natural selection), are seen as designed objects and are therefore interpretable from the design stance. So, for Dennett design does not require a designer!

Not only hearts and legs can be seen as designed objects, *minds* can be seen as being designed too. In fact, Dennett believes that the question of what a mind is, is best explained by appeal to the design of minds. Whereas artifacts are usually understood in terms of minds, Dennett turns the table and says that minds have to be understood as kind-of artifacts, objects optimally "designed" by Mother Nature. This biological take on issues of the mind is confusing at times, but it is interesting to explore because it drives a wedge into a few central debates in philosophy, in particular the mind-body problem in the philosophy of mind and action.

Dennett's wedge makes the mind-body problem not a problem about physics on the one hand, and psychology on the other. Physics is hardly helpful in understanding meaningful behavior. The mental therefore should not be directly connected to the physical; instead, the mental, and goal-directed behavior, should be understood in terms of biology, especially the theory of evolution. The underlying – simplified – intuition behind this idea is that at some point of evolutionary history, physical parts started to evolve into more and more complex biological organisms, that gradually started to develop a mind. The mystery of intentional behavior, then, must be sought somewhere in the process of biological evolution, and not in physics. The design stance plays an important role in the process of understanding what minds are – and why the intentional stance often works so well.

The thought that minds can be explained in terms of their design has made Dennett a prime adversary of the theorists who believe in original intentionality *versus* derived intentionality. From Dennett's perspective, derivitably intentional items such as sentences and technical artifacts may be derived from human minds, but human minds are just as derived: from our design as biological organisms. So the appeal to original intentions is, according to Dennett, a hopeless and spurious solution.

If human minds cannot explain technical artifacts, what can? Dennett's solution is to formulate a notion of design that does not appeal to intentionality but to optimality. I therefore call his solution the *optimality account* of technical artifacts (as opposed to the intentionalistic account of technical artifacts).

The distinction between intentionalism and optimality can be illustrated rather simply by means of an example. Consider a knife. What does make this steel object with a sharp edge a *knife*, instead of just a steel object with a sharp edge? Is it a knife because it was created with that purpose in mind (intentionalism)? Or, rather, is it a knife because it just cuts very well (optimality)? Dennett chooses the second answer: the function of a thing, be it biological or technical, is always what it is best able to do. Perhaps confusingly, Dennett sometimes puts this by saying that designs have to be seen *as if* they were designed by an *ideal* designer.

The optimality account does take seriously the idea that a pure physical description of technical artifacts is not enough to fully understand them. Optimality, after all, is a highly normative notion, and there are no norms in physics. But by avoiding reference to intentionality, it is an interesting attempt to prevent problems that arise from intentionalistic accounts of technical functions.

Another interesting feature of Dennett's theory is that *both* technical and biological functions can be understood by means of the optimality account. This *generic* normative notion of function is a thought-provoking minority view that has gained opposition from two sides. As such, he has to oppose two traditions in philosophy. Opposition comes on the one hand from those who want a notion of function devoid of references to purposes, designers, normativity or any other suspicious teleological element (traditional naturalists). On the other hand, theorists of function that want an intentionalistic notion of function are equally opposed to Dennett's solution: for how can we understand design without a (intrinsically intentional) designer?

Whether Dennett can steer a naturalistic course between a non-intentional, yet normative notion of function, is to be seen in the remainder of this dissertation. But the attempt is interesting enough to explore in detail. Dennett's way of breaking through age-old discussions about the mind, by way of the design stance, is interesting by itself, but certainly for the philosophy of technology too. I started this introduction with the philosophical distinction between body and mind, or between the space of causes and the space of reasons, and the challenge for philosophers of technology to explain how technological artifacts relate to both spaces. Given the claim that technical artifacts do not fit nicely into either category it makes sense to explore whether Dennett's wedge, that opens up a space of biological design, helps us to understand technical design (working according to the same logic as biodesign) better as well.

Dennett's 'relaxed' and pragmatic attitude towards the "hard" philosophical problems may furthermore be very stimulating and productive for a relatively new field like the Philosophy of Technology, where research is still largely in an explorative phase.

Remarkably, however, despite the apparent importance of Dennett's work for the philosophy of technology, Dennett is rarely mentioned in the literature in the philosophy of technology. Carl Mitcham's massive overview of the philosophy of technology does not even mention Dennett. And even though his theory of mind (intentional systems theory, theory of consciousness) has gained a lot of attention in philosophy, the implications of his theory of mind for his theory of artifacts has never been investigated in detail. For a field that is so much related to the philosophy of mind, even forms a part of it, the field shows a considerable lacuna.

Mitcham argues[3] that there are probably two reasons why Dennett is so invisible in the philosophy of technology. Firstly, Dennett has never sought to enter the debates in the philosophy of technology. The field apparently still

[3] In personal communication.

suffers from its reputation of being a new-born field, largely dominated by quasi-philosophical ideas of engineers. Secondly, vice versa, philosophers of technology see him, primarily, as a philosopher of mind and evolution. Indeed, Dennett's theory is a package deal, and his ideas about technology are only a small part of a much larger theory. But this is exactly a feature that I find so valuable of Dennett's work for the Philosophy of Technology, as it may help pull the Philosophy of Technology into established fields in philosophy.

Dennett's design stance, on the other hand, is not tool-made for the Philosophy of Technology. On the contrary, it is designed to complete Dennett's philosophy of mind. Dennett's claim that design is always optimal either in case of biological design or technical design, has gained much more attention (and criticism) in its application to biology, than in its application to technology. The obvious reason for this is that biology is, or seems to be, the problematic case: speaking of design in nature is very controversial, due to the association of the word 'design' with an intentional designer, and assuming design in nature to be optimal is equally controversial. Technical functions are, by many philosophers, seen as either a hopeless or a trivial notion. Dennett is guilty too: he has introduced the design stance primarily in order to understand *biological phenomena*, including the mind. Cultural phenomena like technology are only derivitably interesting, as quasi biological items, or as the apparent unproblematic case. The implications of Dennett's theory of biofunctions for a theory of technical *artifact* functions, have never been investigated thoroughly and have not been worked out in detail by Dennett. As a result, the design stance as applied to technical artifacts needs to be further investigated and evaluated, and has to be constructed from the bits and pieces that are scattered throughout his work (an exception is Dennett: 1990).

I want to emphasize that even though the design stance as applied to technology or artifacts has had little attention (both from Dennett and his critics/admirers), it is a *crucial* element in Dennett's theory that deserves much more attention and a better look than it has so far gotten. The design stance *should* work for technical functions, on pain of having to drop the claim that technical functions and biological functions can be understood in the same terms; and this claim is crucial to Dennett's theory. Making a special case of technology would undermine his key claim that all intentional and cultural phenomena, in the end, can be biologically explained. But this only succeeds if Dennett can show that a notion of technical function without reference to a designer can be made to work. So there is something at stake not only for the Philosophy of Technology, but for Dennett as well.

1.1. Specification of the main question

The elemental question in the philosophy of technology, and certainly within the Dual Nature project, is how artifacts and artifact functions have to be conceptualized. In this thesis I will review the prospects of Dennett's theory of function ascription (the theory of the design stance) and address the question whether Dennett's method of the stances provides an adequate framework for conceptualizing of and attributing functions to artifacts. The general question reads:

> *Does Dennett's method of the stances provide an adequate framework for the conceptualization and attribution of functions to artifacts?*

'Adequate' has to be specified. I will formulate two *internal* standards, empirical correctness and methodological utility and evaluate the theory on the basis of that. Furthermore, as will become clear in chapter 3, a conceptualization is nothing more or less than attribution for Dennett, we can reduce the question of conceptualization to a question of attribution. This turns the general question into a more specific one:

> *Does Dennett's method of the stances,* according to standards inherent in his kind of position, *provide an adequate framework for the attribution of functions to artifacts?*

I will evaluate Dennett's method of the stances as applied to technical artifacts on the basis of these two standards.

A few qualifications are in order. Firstly, I will be doing a lot of interpretation of Dennett's theory.

Interpretation will thus be a main part of the thesis. Interpretation is not always seen as a proper form of investigation in science, and interpreting the work of a *living* person may strike as redundant. One might argue that the philosopher is, after all, in the best position to answer questions about what he means. Conveniently, Dennett himself disagrees with this position (I will discuss this claim later in this dissertation): an interpretation may provide a better understanding than the comments of an author on his own work. It may be more systematic, may reinterpret a view from a different angle, etc. I should add that my interpretation is not the kind of historical interpretation of the kind where we enter the head of the artist, and try to find out what he *really* thought when he wrote what he wrote. I will use a more theoretical-systematic, charitable kind of interpretation, specifying what I think are the basic ideas behind the theory, fixing those, and interpreting the rest of the theory on the basis of these basic ideas.

My philosophical interpretation discharges into an internal critique of Dennett's theory of artifact function. On the basis of two standards that I take to be fundamental to Dennett's views, I evaluate his theory of artifact function – more in particular *the optimality principle* that underlies Dennett's account of the design stance.

Note that I prefer to write 'Dennett' in stead of 'Dennettian' throughout the thesis, but the conclusions I draw should hold for everyone holding a similar position. The two standards are basic to Dennett's view, but they are not limited to it. They should be or are in fact embraced by many philosophers, certainly those working in the more naturalistic tradition. As such, this thesis has relevance beyond Dennett's own ideas only. The implications of the two standards hold in principle for everyone that values them. (Whether the position is relevantly similar, can be determined on the basis of the two general standards and four theses I define for a Dennettian theory of ascription in chapter 5).

As mentioned, a theory of artifacts and technology cannot be seen apart from other issues, especially issues in the philosophy of mind. Established philosophical approaches are not free to take any position they like with regard to questions about technology. Dennett's general project, being a well worked out and established approach in an established philosophical field, is worth exploring further as a position in the Philosophy of Technology.

I deliberately choose not to evaluate Dennett's theory on the basis of standards derived from the Philosophy of Technology. The field is not mature enough yet to provide such standards, and if there are standards it is exactly at stake whether they hold in the light of debates in established fields like the Philosophy of Mind. Dennett's theory of the design stance is only going to be valuable as 'established input' for the Philosophy of Technology when we can be sure that it at least holds within his own theory of mind.

My approach will be 'top-down', starting with Dennett's most general philosophical view, through the philosophy of mind and the theory of the intentional stance, until I can derive from it, as systematic as possible, his views on functions. This is necessary, because this is the actual way in which Dennett has developed his idea, *and* because it helps me constrain my interpretation of the design stance. As a consequence, it will only be in chapter 5 that the design stance makes its entrance.

I have chosen to set this thesis up as an internal critique for another reason. We can distinguish between external and internal critiques. In an external critique, the theory is attacked from the outside with external standards. Obviously, the discussion can then only be settled on the basis of a very fundamental debate about these standards. Debates, then, are usually about very fundamental issues, for instance about what a mind is, what a correct conceptualization is, what consciousness is, etc. (in chapter 2 I will globally review a number of fundamental debates). Dennett is, as a matter of fact, often externally criticized. The philosophical game then often comes down on picking your external standards cleverly, and rejecting the theory with them.

External critiques are important because they force us to review the basic standards that determine the outcome of the discussion. But they often do not really hurt the theory, because external standards can in the end be easily dismissed *as being external*. A theory can be rejected *only* if it is certain that the external standards that are used are beyond doubt. And this is rarely the case. An internal critique is much stronger. Once you can show that a theory is internally inconsistent, or that the theory does not fulfill standards that are basic to the paradigm, the theory is in direct trouble.

My analysis of Dennett's theory of technical artifacts is directly driven by research questions in the Philosophy of Technology, but this dissertation is only a first step to answering the question whether Dennett's paradigm can be successfully imported into this field, let alone whether it can solve its philosophical problems. The goal of this dissertation is to get clear *what* Dennett says about artifacts (a question that turns out to be a lot more complex than one might expect), and to internally evaluate it. My first concern, then, is with Dennett. Only when we get to the point of getting enough clarity about the position, a position that should be internally stable and convincing, is it useful to evaluate its virtues for the Philosophy of Technology. I do not want to give away too much of the conclusion of this thesis, but in order to prevent disappointment: that point will not be reached in this dissertation.

1.2. Short overview of the thesis

After giving a global overview of the relevant debates in the philosophy of mind, or the philosophy of attitudes more specifically (chapter 2), I lay out Dennett's general philosophical project and derive the two standards – empirical correctness and methodological utility - from it (chapter 3). After a short and rather technical excursus into the philosophy of science (chapter 4), only indirectly relevant for the main argument, I proceed to explain in detail Dennett's theory of artifact interpretation: the optimality account (chapter 5). Chapter 6 evaluates the optimality account on its empirical correctness (first standard), chapter 7 on its methodological utility (second standard). Results are evaluated in the conclusion in 8.

1.3. A matter of courtesy

It has become common practice in philosophy to use 'she's and 'her's when referring to persons. I am a post-feminist and take no offense when people use 'he', 'him' or 'his'. 'Person', after all, is masculine. But I certainly do appreciate the gesture, and shall return the courtesy in this thesis by using 'he', 'him' and 'his' throughout the thesis.

2 Attitude philosophy

In this chapter I shortly introduce the philosophy of attitudes in terms of three important kinds of orienting distinctions that help identify the relevant philosophical positions in the debate about attitudes and specify the basic terminology.

The philosophical debate about attitudes is important, because it penetrates many others, most notably the discussion about functions – and thus artifacts. Dennett, as well as many other analytic philosophers, derive their theory of function from their theory of attitudes. The most common view is that we have to understand artifacts as functional items; and thus that (artifactual) functions have to be understood at least partly in terms of human intentions and attitudes. The mere physics of an artifact do not make it an artifact. It is an artifact – as a functional item - also because of our system of values, our intentions, and our thoughts about it. (see, e.g. Searle: 1996, Baker: 1995, Dipert: 1993, McLaughlin: 2001). In addition, we individuate an artifact (this is a car, this is a sock) on the basis of the its function. Hence, we need to know something about attitudes to understand artifacts.

We ascribe beliefs, desires, intentions, wishes, fears and many other kinds of attitudes to a great number of entities. The Egyptian supermarket owner around the corner hopes to make more profit (so he sells freshly squeezed orange juice). I personally want to get a Ph.D. soon (so I sit behind my computer and write these lines even though the sun is shining). The monkeys in the Artis Zoo are frustrated (so they scream a lot, and make crazy jumps). Cleo the kitten knows I find her irresistibly cute when she lies on her back playing with the mouse (so she uses this knowledge to trick me into giving her a treat). My computer hates it when I log into the Eindhoven server system (so it blocks my regular e-mail whenever I try anyway). The fig plant on my balcony craves for some water (so it drops its leaves – as an act of revolt,... or is it just sad?). And that mean loose tile over there on the street is just *waiting* to let me trip over it...

Obviously, many of these ascriptions are merely metaphorical. We all know that fig plants and computers have no emotions, and that kittens have no knowledge about other's states of minds. And you must be pretty paranoid if you think that tiles intend to harm you. No: only human beings have the full capacity to think, hope, hate, deceive, revolt, and to be sad. Or... so it *seems*.

The philosophical debate about the mind can be roughly divided in two topics: intentional attitudes, and qualia. Intentional attitudes are the psychological states we ascribe to people in everyday life. They are often called 'propositional attitudes', 'intentional states' or just 'attitudes'[4]. 'Qualia' refers to the qualitative aspects of phenomenological experience. Feeling pain, for instance, or seeing red. Intentional attitudes are *about* something, they represent[5]. They are, in philosophical jargon, 'contentful'. Amongst the intentional attitudes are such events as believing, knowing, desiring, hoping, and wishing, but also dreaming, remembering, imagining etc. The habit of explaining behavior in terms of attitudes is sometimes called 'folk psychology'. Although I will use the term 'folk psychology' occasionally when I talk about the habit of theorizing about the mind in terms of attitudes, I generally prefer to use 'attitudes' because 'folk psychology' is often associated with a research program that is much more specific, and already biased towards a certain position in the debate about attitudes (I will explain this in chapter 6).

It is a fact that we ascribe attitudes to many things in daily life, and that we predict and explain the behavior of these things (their 'actions'[6]) on the basis of these ascriptions. Sometimes we take such ascriptions quite literally, for instance when we ascribe attitudes to human beings. Attitude ascription allows us to make remarkably successful predictions of the actions of other human beings. In fact, most if not all social life would be impossible without using attitude-talk. It is, for instance, hard to see how we would make promises, appointments, or

[4] It the thesis, I generally use the term 'attitudes', following Lynne Baker (Baker: 1995).

[5] The distinction between intentional states and qualia is an artificial and theoretical division; intentional states may have qualitative aspects; consciously being in a certain psychological state, for instance. Dennett believes that intentional states are the primitive case (see p.27) so the distinction is adequate for my purposes.

[6] For the moment, I will assume that any behavioral prediction that is based on the ascription of intentional attitudes yields an action explanation; that is, that 'action' is behavior phrased in intentional language. So if I ascribe to you the belief that my cat Cleo will purr if you stroke her, and the desire that Cleo purrs, I will predict that you will take the action 'stroking Cleo'. If I were a very brilliant neurophysiologist that knows your brain through and through or a Laplacian demon that could predict every coming causal event on the basis of the laws of nature, I could predict that your *hand* will in a few seconds move towards *Cleo* (*'s mark very sloppy Laplacian language use). But I would, in that case, not be describing an action of yours.

have any other form of coordination without them. The explanation for this may seem quite simple at first: human beings have minds (thoughts, wishes, hopes), *and* we are excellent mind-readers, and thus are able to “read”, at least by approximation, the thoughts and wishes of fellow human beings. Thus, if I predict that you will be at your office at 4 p.m., because I believe you will keep the promise that you made to me a few weeks ago, and if my prediction turns out to be true, part of the explanation of the predictive success lies in the fact that I “got your mind right”. You *are* indeed a promise-keeper. If you are a realist about intentional attitudes, you believe that some such explanation must be correct (an intentional realist thinks that an intentional attitude is real, that it exists in the world, usually in the heads of people, and that attitudes affect states and events in the world).

Suppose for now that human beings do indeed have attitudes. How about attitudes of non-humans? Do they really exist, like we think human attitudes exist? We use the same attitude-talk for non-humans: animals, other organisms, and artifacts. Think of the bored monkeys in the Artis Zoo, the fig-plant’s revolt, and the computer’s hate against the TU/e system. We ascribe attitudes even to organizations (“the Dutch government hopes to regain the trust of the people”) and inanimate things (“the moon follows me wherever I go”, “my car wants to drive to the left all the time”).

In many cases, human and non-human alike, such intentional ascriptions are very successful: we are able to explain and predict the behavior of the entities at stake. It is often thought that intentional ascriptions are technically only correct when ascribed to human beings. In most other cases, it must be mere metaphor; convenient but sloppy language use. Indeed, many philosophers think that the ascription of attitudes to non-humans (except, perhaps, the higher primates) is literally false.

But on what basis are we going to determine that human beings really *do* have attitudes, but (certain) non-humans do not? If we take attitude ascriptions to human beings seriously, why not those to non-humans? It cannot be that attitude ascription to non-humans is not explanatory or predictive. The Cleo example was exactly meant to show that. And even if we find some characteristic of human beings that discerns them from other beings (say, their having language), there still remains this one question: Attitude explanations of animal behavior *are* powerful – much more powerful than we would expect from a "mere metaphor". There must be *some* truth to such attitude ascriptions. What kind of truth can that be?

Artifacts play a special role in the debate. Might *they* have attitudes? A widely held view is that *human beings* have beliefs, desires, hopes and fears, and that perhaps *some animals* have them (in some sense), but that *artifacts*, especially rather simple ones, are a special case. You may perhaps grant that it is in principle possible to create an artifact that would have beliefs, but claim that we have in fact not yet created such an artifact. Or you may think that even the most advanced robot could not have proper beliefs, because the "beliefs" it has are necessarily derived from us: *we* "implant" them[7]. Nevertheless: we do in fact ascribe attitudes to artifacts, and quite successfully so (just think about the conversations you had lately with your computer). And it is not at all clear why such ascription would be wrong in the case of artifacts, yet justified in the case of human beings. On the basis of what criterion exactly are we going to determine that?

Apart from the question whether we can make a proper and clear distinction between attitude ascription to human beings and attitude ascription to non-humans, and sometimes driven by it, many philosophers doubt whether even human beings really have attitudes. The predictive and explanatory force of attitude explanations is compelling, but this may require a different explanation than the existence of beliefs – as we will see.

So we have a certain kind of language use, an intentionalistic one, that is pervasive in human life, that is indispensable for many social interactions, and that we even apply to non-human entities. The philosophical debate about this intentionalistic language has gained a life of its own. It has resulted in a rich and complex philosophical research program within the philosophy of mind, with several angles and ways to approach the issues, and involves deep and broad questions not only in the philosophy of mind, but also in the philosophy of science, ontology, epistemology and even the philosophy of language.

As a result, there is hardly a fixed philosophical meaning of the word 'attitudes'. Take 'belief' for instance. A philosopher may be talking about the

[7] The options are, obviously, not exhaustive. Daniel Dennett, for instance, would grant even simple artifacts beliefs (Dennett: 1987b). But the two options above are the most widely held opinions about the ascription of attitudes to artifacts.

ascription of a belief, about a belief itself, about what we mean by belief, or what it actually is. He may be talking about the (non)existence of beliefs in brains, or in behavior, or about the way we use the word 'belief' in everyday language. The philosophy of attitudes is, as a result, really about several (related) topics, and not about one. For someone unfamiliar with the philosophy of attitudes, a random paper in the philosophy of attitudes will probably be rather alienating.

I do not aim to clear up the confusion in this chapter, nor can I do justice to all the intricacies in the debate. What I aim to do in this chapter is to plough my way through the philosophy of attitudes so that I can explain Dennett's position in the debate, and explain certain choices Dennett makes later. This chapter thus provides the terminology that I will need to make Dennett's ideas about the mind, artifacts, and design clear. I will make a number of rough basic distinctions that we already briefly encountered on the first pages of this chapter. This should help to "fix" the meaning of basic terminology I am going to use, and will help clarify some issues that will turn up later when I discuss the main paradigm of the dissertation: the interpretationist approach of Daniel Dennett (chapter 3).

The rest of the chapter is structured as follows. I will first shortly discuss some (alleged) characteristics of attitudes. They are usually seen as being perspectival and subject to rational evaluation, and can be divided roughly in belief-like kinds of attitudes, desire-like kinds of attitudes, and intention-like attitudes. I will also deal shortly with the claim that attitudes are essentially tied up with a first person perspective (2.1).

Then I will proceed in dealing with the main distinctions. The *first* distinction I will discuss is that between **attitude ascription** and **attitudes themselves** (2.2). The *second* distinction is between **normative** and **descriptive** approaches to attitude ascription (2.3). The *third* distinction I will make is between what I call **folk approaches** and **scientistic approaches** to attitudes (2.4).

These three distinctions must be seen as guiding idealizations – tools to work my way through the philosophy of Dennett. The main purpose is to clarify Dennett's position. The distinctions point at competing ways to approach attitudes. But for Dennett, as we will see in the next chapter, all these distinctions collapse.

2.1. Characteristics of attitudes and attitude-explanations

An important characteristic of attitudes, due to their intentionality, is that they reflect the perspective of the one having the attitude. Philosophers call this the perspectival nature of intentional attitudes. Attitudes are about something, about objects, or states of affairs in the world (or reality, if you like), but they are not necessarily true about the world. Attitude explanations explain behavior *from the inside*, and therefore take into account the way the world is *according to the agent*. This is important, because it is exactly this characteristic of attitudes that makes the explanation in terms of attitudes so powerful. For instance, my belief that the train to Eindhoven will leave at 11.06 am, explains why I leave in a hurry at 10.45 am to the train station. That the train is in fact cancelled is irrelevant for the explanation of my going to the train station. Another example: the reason that I put salt in my coffee, is that I thought it was sugar. On the basis of attributing to me these (false) beliefs in these cases, by taking *my* perspective on the world, you could readily have predicted: (1) my surprise and anger when I entered the train station, and (2) the strange look on my face when I drank the coffee.

Another characteristic of attitude ascription is that action explanations in terms of attitudes involve some kind of (practical) rationality. From the beliefs and desires we attribute to an agent, we make a prediction what the action will be - or we explain the action in terms of the attitudes we ascribe to the agent. But such conclusions can only be drawn when we presuppose that the agent is rational.

Consider an simplified attitude-explanation. Let's say that Sue has the following attitudes (and, for the sake of the argument, no more):

(1) Sue wants to eat bread or meat
(2) Sue believes that bread is bought at the bakery
(3) Sue believes that meat is bought at the butcher
(4) Sue believes that the bakery is open
(5) Sue believes that the butcher is closed

What follows? What is Sue going to do? Go to the bakery of course. If she were to go to the butcher ... she would be irrational. Our prediction that Sue will go to the bakery presupposes *that* she is rational. That is to say, we need an extra assumption that Sue is rational and will therefore take an action that will satisfy her desire for bread or meat.

Rationality also means that the beliefs of an agent are internally consistent, and that he believes the implications of his beliefs.

The kind of rationality we need for intentional explanation is quite minimal. Being rational is not the same as being moral, nor as being particularly smart. Rationality, in this context, is the sometimes quite simple practical deduction from a set of beliefs and desires (or other attitudes), sometimes called means-ends rationality. But it is indispensable. When our intentional predictions fail, we first try to revise the premises in the light of attributed rationality. For instance, if it turned out that Sue went to the butcher after all, we may question whether she indeed believed that the butcher was closed (wrong attribution of belief), that she may have had other desires (wanting to check opening times of the butcher), and so forth. So in an ordinary attitude-explanation, we *keep assuming* that she is rational, and adapt the explanation to that assumption.

How exactly rationality comes into an action explanation, and how rational we have to take an agent to be is under heavy debate (see, e.g. Millgram: 2001). What matters for this thesis is that rationality plays an indispensable role in action explanations. In chapter 3, I shall discuss Dennett's interpretationist explanation of this fact.

Attitudes are usually divided into two (or three) categories: beliefs, desires, and intentions. Beliefs and desires are most discussed. Intentional explanation is thus sometimes called *belief/desire* psychology (e.g. Fodor: 1987). Philosophers also talk about the 'belief-desire'-model. The reason for this is *not* that philosophers think that beliefs and desires are the most important attitudes. Terms like 'belief' and 'desire' belong to the philosophers' jargon. The term 'belief' stands for a kind of informational state, and includes attitudes like thinking, guessing, perceiving, remembering etc. They are attitudes that represent how the world is to the agent. "Desire" stands for a pro-attitude, like wanting, longing and hoping (cf. Davidson: 1980). They are attitudes that represent how the agents would *like* the world to be. This distinction is sometimes formulated in terms of their having a different 'direction of fit' (Anscombe: 1957). Beliefs have a world-to-mind direction of fit, that is, the content of a belief is supposed to fit the world (my belief is false when it does not fit the world). Desires have a mind-to-world direction of fit, that is, the world is supposed to fit the content of desires (my desire is unsatisfied if the world does not fit it). Apart from beliefs and desires there are intentions. Intentions are special because they are a kind of intermediate attitude – intermediate between having certain beliefs and desires, and performing an actual action. Beliefs and desires are then said to "cause" actions *through* intentions (see in particular Bratman: 1987). Note that intentions require belief/desire attitudes. And note that the term 'intentional states' is (perhaps as a result) often used to cover *all* attitude states.

A hot topic in the attitude debate concerns the question whether having a belief and ascribing a belief requires having a first person perspective: a perspective from the agent having the attitude. What exactly a first person perspective is, is not so clear. A *weak version* of first person perspectivism may be the view that a first person perspective just describes how the world is from the point of the view of the agent (what is called the perspectival nature of attitudes, or a perspectival attitude, or a weak first person perspective, cf. Baker: 2000, 61). You can describe the point of view of the agent from an outside perspective, for instance, as we do with animals.

Usually, however, a stronger version is suggested. Unfortunately, it is not so clear what exactly this strong version of the first person perspective is supposed to be. There are many versions of this view around and 'first person perspective' has become a container-term, a buzz word even. I will not attempt to clean up the mess. The first person perspective is often called the perspective of the "I", your own perspective, your subjective, conscious, or internal point of view (cf. Baker: 2000, especially chapter 3, Baker: 1998, cf. Shoemaker: 1996). It is often connected to qualitative aspects of the mind (qualia). The claim of the strong first person perspectivist is that having an attitude has to do with "having a self", being able to conceptualize oneself as oneself, with a certain kind of "privileged access" perhaps. The crucial claim made by the strong first person perspectivist is that from the first person perspective, we can gain insight about the mind (or the content of an attitude) that *cannot be had from* the outside, or the *third person perspective*.

The third person perspective is the perspective of the "he" (or she, or it). It is therefore necessarily "external", an outside perspective. Naturalists are typically third person perspectivists. The reason for this is that naturalist theories rely on scientific methods, and science is based on the third person perspective (a scientist could never use introspective data as validating proof for his theory – cf. Dennett: 1991a, 70-71). The naturalist thus can embrace only weak first person perspectivism.

An important question is whether or not we can understand attitudes from the third person perspective. Naturalists (but also some non-naturalists) like to think that they can, but many philosophers believe that this will turn out to be impossible, because attitudes have first personal elements that can never be captured from the third person perspective. The third person perspectivist can reply in two ways. First, he can show that having an attitude does not require having a strong first person perspective. Second, he can try to show that although the first person perspective is important to understand some aspects of intentional attitudes, the valuable elements of it can be (and have to be) perfectly understood and explained from the third person perspective.

Those who think that an attitude is always an attitude-with-a-first-person-perspective will usually grant only human beings attitudes, for only human beings have seem to have first person perspectives[8]. But there are other reasons to think that only human beings have attitudes. (These reasons are available both for the first person perspectivist, as well as the third person perspectivist). For instance, one could argue that an attitude always requires a concept of that attitude, e.g. a belief always requires the concept of a belief, which requires human language. Or one could claim that having a belief requires being able to ascribe a belief, i.e. having a theory of mind (cf. Davidson: 2001; Davidson: 1994). Some such story, usually with the point that a belief has to be "conscious", "reflective", "phenomenal" or "entertained", or something like that, will effectively exclude all or most non-human beings from the domain of believers. In chapter 7, in which I deal with animal beliefs, I will give some counter-arguments against such claims that set high standards to what counts as a proper attitude.

2.2. Attitude ascription and attitudes themselves

Let me quickly go through the debate about attitudes by means of three distinctions. The first distinction I want to introduce is between two subject matters:

[8] The first person perspectivist may grant other animals a first person perspective, depending on how he defines what a first person perspective is. Some first person perspectivist may think that dolphins and chimpanzees have first person perspectives if it turns out that they have some form of self-consciousness, or if they have the capacity to have higher-order attitudes, e.g. when they pass the mirror recognition test, and the false belief test. More about higher order attitudes in chapters 5 and 6.

attitude ascription and *attitudes themselves* (as I will call them). This distinction may seem rather straightforward. Attitude ascription is a judgment about a phenomenon in the world having certain properties or characteristics – just like other judgments we make. If we ascribe an attitude we make a statement about an agent that has the property of having a certain belief (or desire). This statement can be right or wrong. There is no fundamental difference between, say, judging that "that flower is red", "the apple falls from the tree", and "that person wants to cross the street". An example. Margaret ascribes to George the belief that he wants to cross the street: "George wants to cross the street". Ascription, then, is about *Margaret saying* something about the world, namely, *that George wants to cross the street.* The attitudes themselves are about *George,* or his attitude "*I want to cross the street*" (or "I don't want to cross the street"– Margaret may be mistaken about what George wants).

Dennett has phrased the distinction between attitude ascription and attitudes themselves in terms of 'folk psychology as craft' and 'folk psychology as ideology'. The craft is about the actual ascription of attitudes, the way we actually do it, learned to do it, and explain and predict behavior with it. The ideology is "what ... the craft [was] all about" (Dennett: 1991b, 137), that is to say, what attitudes seem to be when we ascribe them to other people and ourselves. In other words, our daily habit of ascribing attitudes (attitude ascription, the craft) has lead to an ideology of what attitudes themselves are (if they "are" at all). According to Dennett, there is "room for false ideology" (135).

Let me clarify by comparing it with 'folk physics'. Folk physics helps us predict and explain the behavior of inanimate bodies (that the sun rises, that a ball falls with a certain speed etc.). This is the craft. We can study the craft of folk physics as we can study the craft of folk psychology (as an anthropological research). Or we can try to systematize it in science. But we cannot take the ideology of it - 'that the sun rises' – at face value. The everyday theory of how the physical world works may be (in fact: is) wrong (136). Similarly, we cannot simply assume that our daily folk psychology has it right. There has to be further argumentation as to what the relation between the craft and the ideology is, and philosophers disagree about it. Note that it is not so much the question whether a particular ascription is right or wrong – that Mary may be wrong about George. The issue is whether the whole *theory* of attitudes, that underlies our daily ascriptions, is largely correct. Intentional realists (e.g. Fodor) claim that our daily folk psychology, as a theory, is in fact largely correct. Dennett is not so optimistic

– folk psychology, just like folk physics, has it wrong at certain crucial points - hence his choice for the word 'ideology'. The craft is fine, but we need to be very careful in taking the ontology behind it ('attitudes themselves') literally (I will get back to this issue in 3.6).

Straightforward as the distinction between attitude ascription and attitudes themselves may seem, in philosophical and empirical practice the two are easily mixed up. The main reason for that is that when I ascribe a belief to someone, I am *having* a belief, namely: that this someone has this particular belief. Attitude ascription is a mental process, and a rather complicated one at that. So when we talk about attitude ascription, we could be talking about either the mind of the one being ascribed an attitude, or about the mind of the attitude ascriber. For example, if I say "Peter believes that they do not make vacuum cleaners as they used to be", this may tell us something about Peter's mind (his beliefs), or about mine (my apparent belief that Peter believes that...). Many philosophers are more interested in the former, more particularly in the question whether such an attribution is *true*, and how we determine that. Cognitive psychologists (and other empirical researchers of the 'ascribing mind') focus mainly on 'my' beliefs, i.e. the mind of the ascriber. For instance, they try to determine at what age, and under what circumstances human beings are able to ascribe attitudes to other beings; that is, when they have a *theory of mind* (I will discuss details of this and related research later, in chapter 6).

So far so good. Why do these two subject matters get so easily mixed up? Let me give a few examples. Some philosophers think that my ascription of Peter's belief that 'they do not make vacuum cleaners as they used to do', is true only if Peter is able to ascribe attitudes to other people, i.e. when he has a theory of mind (cf. Davidson: 2001, Davidson: 1994). In that case, we can only speak of a real attitude when the one having the attitude (being an attitude ascribed to) can also ascribe an attitude.

Another example. Philosophers that analyze the concept of an attitude, say belief, will typically try to construct a set of criteria that specify both the conditions of having a belief and that of justifiably ascribing a belief: we are only justified in ascribing a belief when there is one. The descriptive set of facts about attitudes themselves has to be aligned to the normative set of criteria that specify attitude ascription. This is a perfectly legitimate philosophical enterprise, of course, but as a result the distinction between attitude ascription and attitudes themselves gets out of sight (or is not even made).

Last example. Some philosophers, like Dennett, think that there is in fact no distinction between actual beliefs, and belief ascriptions. Or rather, such philosophers think that there is *nothing but* belief ascription, and that actual beliefs are just what turns up in an ascription (or, perhaps, a *"good"* ascription). I will explain and discuss this thesis of Dennett in detail in the next chapter – what matters now is that there are many ways to blur, or ignore, the distinction between attitudes and attitude ascription.

To some extent the distinction between attitude ascription and attitudes themselves boils down to the famous distinction between epistemology and ontology. The distinction between epistemology and ontology is a real philosopher's gadget. It runs through the whole history of philosophy, it is directly addressed or it turns up in several disguises[9]. Epistemology is about what we know and about knowledge in general: how we know, proper ways of knowing, and what knowledge is. Ontology is about 'what there is', in "the real world". Epistemology is about the knowing subject, typically us, human beings. Ontology is about what there is independent from us. Philosophers are also interested in their relation: how can we make sure that our knowledge of things is about the things themselves (and not just ... what we *know*). The common idea is that ontology does not care about epistemology, but that epistemology does care about ontology. Applied to attitude ascription, we could say that ascribing attitudes to things (knowing, or thinking, that things have attitudes) is an epistemological fact. Whether or not there really *are* attitudes, is an ontological question.

The distinction between epistemology (ascribing attitudes) and ontology (attitudes themselves) may seem clear at first, but there are two important sources of confusion. First, attitude ascription is not only epistemological, but has an "ontological" side to it as well. That is to say, an attitude ascriber ("the knowing subject") is part of the real world and the process of attitude ascription and the conditions under which this is possible can be studied and described like anything else. We can try to understand how subjects are able to ascribe attitudes, to name but one example. Second, some philosophers think that the ontology of the attitudes – or even the ontology of *everything* – boils down to their epistemology. They claim is that if you have the epistemology of attitudes in place, there is nothing more to say about ontology: the world consists of every-

[9] In the history of philosophy, 'ontology' is often called 'metaphysics'.

thing, and nothing more, than what we could possibly know about it. Dennett holds a mild version of this position (see also 3.6).

Let me close this section by adding to the distinction 'attitude ascription' and 'attitudes themselves' a third category: attitude methodology. This will turn out to be very convenient in the interpretation of Dennett's theory later on. According to Dennett there are two ways to look at attitude ascription: as an everyday explanation of behavior in terms of beliefs and desires (as we have discussed), and as a methodology for a number of sciences (notably evolutionary biology, but also the social sciences). The scientific methodology is an idealized version of the everyday version. Idealized in two senses: in the sense that the methodology is regarded the better version of everyday folk psychology, and in the sense that the methodology eliminates certain unhelpful features of everyday folk psychology (for instance, the idea that attitudes are concrete items in the head) and overemphasizes others (for instance: rationality).

According to Dennett, ascription precedes attitudes themselves. An attitude is what turns up in our correct (successful) ascription of the intentional stance. The correct ascription is then provided by the attitude methodology. So, for Dennett, it is impossible to use our everyday ascriptions directly (for instance, by means of introspection, or an analysis of the language) to get to an account of attitudes themselves. I will discuss this claim in chapters 3 and 5.

2.3. Normative/descriptive

I have already mentioned the distinction between normative and descriptive accounts of attitudes. Descriptive accounts, I said, are about how, when and why we ascribe attitudes to certain entities. For instance, we may want to know what parts of the brain are responsible for the ascription of attitudes, how we learn to do it, and what the biological and cultural function of it is. We may also want to know if other animals, or robots, are able to do it, and if ascribing attitudes is cross-cultural. Another, different, descriptive research program studies the referents of attitude ascriptions: attitudes themselves. If human beings indeed have them, as we usually take for granted, where are they then located? In the brain? In behavior? Do animals have them? And artifacts? And rivers? Organizations?

The *normative* part of the debate about attitudes is about *whether*, *when* and *why* this piece of language is *legitimate*. One may ask, for instance, whether attitude talk is strictly incorrect when applied to non-humans. On what basis do we determine that?

The two tasks – descriptive and normative - cannot be easily separated. For instance, if you think that folk psychology is a theory that describes items that correspond, more or less, to internal (brain) states (descriptive), you may find it rather easy to justify its use in cases involving these (brain) states (normative). (Compare: my claim that there is a cow next to the tree is justified if and only if there is a cow next to the tree).

The philosophy of attitudes is usually regarded to be about normative questions about attitude ascriptions, that is to say, about the question when attitude ascriptions are correct. The theory may be built on descriptive data, but the interesting question is whether or not attitude ascription is strictly correct. Naturalistically minded philosophers will put the emphasis on the descriptive issues, but even the naturalist wants his theory to be 'right'[10].

2.4. Folk attitudes and scientific attitudes

In the introduction of this chapter, I mentioned a few examples of intentional attitude ascription. I ascribed such attitudes as 'thinking', 'being bored', 'hoping', and 'hating' to animals, people and artifacts. I also said that the question whether and when such ascriptions are justified (the *normative* question) can be answered in a number of ways. But do we seek the ordinary meaning of the terms, or go for scientific vindication? Or, analogously, do we seek a folk justification of our attitude ascriptions, or aim for a scientific one? So we may be looking for folk attitudes, or scientific attitudes.

The terminology is mine, and is only meant to tentatively distinguish very roughly two approaches in philosophy; and only in order to help me explain Dennett's approach.

Let me start with folk attitudes. Ascription of attitudes belongs to the domain of everyday language, and everyday explanation. That is why the ascription of attitudes is often called *folk* psychology. If we want to evaluate the ascription of

[10] Some naturalists aim to stick strictly to empirically describable fact, e.g., the naturalistic project ´descriptive epistemology´ in epistemology, a field that is dominated by normative issues. See, most notably, Quine: 1988 and Kornblith: 1988 for an overview of positions.

attitudes in ordinary life (the normative question[11]), it makes sense to do that on the basis of 'folk criteria'; the way we use the terms. But what exactly are folk criteria, what exactly are they supposed to justify, and what can we expect from a "folk justification"?

A possible answer goes something like this. We first want to know what we *mean* exactly when we use folk psychological terms, like 'belief', 'desire', 'knowledge', or 'intention'[12]. Our primary source is how we use the terms in our language, and the corresponding semantic intuitions. In order to know that, we look at exemplary instances of the use of the term in the language, and formulate a consistent definition of the term, that includes the positive instances, and excludes the negative ones. On the basis of this definition, we can then try to determine whether we can apply the term to a new case. If the new case does not fit the definition, we may want to adapt the definition in order to fit the new case, or we may decide that the term is not applicable to this new item. Whether we adapt the definition, or exclude the new item depends on a number of considerations, for instance: whether the definition becomes too broad when we adapt it, whether it forces us to include cases that we clearly do not want to include, whether there are other terms available that would cover a broader range of items, etc.

This approach, let's call it conceptualism, is usually normative in nature. Knobe describes it very well as the position that says that:

> a concept has something to do with the *correct* use of a word. When we offer an account of, e.g., the concept of intentional action, we are offering a standard against which people's ordinary utterances can be judged. If people do not actually use the words "intentional action" in the way specified by the account, we might conclude that they are speaking incorrectly. (Knobe: 2003, 312)

A lot of the decisions that we make in this definition process are eventually based on intuitions about how to use a certain term. Often, philosophers make

[11] Most philosophical investigations of folk attitudes are normative. But not necessarily so. Experimental philosophy, for instance, sticks to mere description of our attitude terminology.

[12] The method is very popular in the philosophy of action. Concrete examples are found in action theory (e.g. "intentional action", see, e.g. Mele: 1997, but Anscombe: 1957 is also a good example), in ethics (e.g. "goodness"), and in epistemology (i.e. analysis of the meaning of 'knowledge').

an appeal to the reader's semantic intuitions in order to defend their proposed definition of a certain term. Our reflective intuitions about the right application of a term are then used to come to an acceptable definition. Someone may challenge the proposed definition by giving counter-intuitive examples[13]. (I call this a holistic definition process, because singular facts are constantly evaluated in relation to each other and to the larger picture they make up.)

Take, for example, lying. What is 'lying'? Certainly it includes something like 'not telling the truth'. But there is more to it. A *mistake* is not a lie (when Ptolemeus said that the earth was flat he was not lying, he was merely making a mistake). What is important for our concept of lying is that we *intentionally* do not tell the truth. So our first definition of lying is 'intentionally not telling the truth'. But there is still a lot of fine-tuning to do. For instance, take the next 'thought experiment'. Mieke tells her friend Sally that Sally's husband is unfaithful, *believing* that this is not true (Mieke is jealous, say). But, unbeknownst to Mieke, Sally's husband is indeed cheating, so technically Mieke told the truth. Did Mieke lie? And there are more questions: is intentionally withholding the truth lying? Can you lie without speaking? Is a pathological liar really a liar? When do we do something intentionally at all? Etcetera.

So, conceptualism tries to find the *meaning* of a certain concept. Often, but not always, the supposition is that although the language is sometimes messy and confused, meanings (or concepts) itself are consistent, and that there are clear and systematic definitions to be given.

Such investigations are not only of theoretical interest. They are also valuable for use in legal and political contexts, as well as in the moral domain. For instance, take the case where we are to decide whether the suspect was responsible for a certain action. We expect a judge to give a judgment based on a notion of responsibility that is repeatable and consistent and we expect that the criteria that he or she uses are applicable to new cases without violating our intuitions too much. In such cases, conceptualism is a very useful and important one. Similarly, the concepts of (intentional) action, knowledge, and so on, can be well worth investigating.

[13] Intuitions, in such cases, count as "data" that have to be explained within the proposed theory, similarly as in science. The more data (intuitions) that can be accommodated, the better the theory. Thought experiments help to get our intuitions sharper, or to decide between conflicting intuitions.

But one may wonder to what extent conceptual analysis of attitudes will give you any insight in the "real nature" of attitudes – for all we did was analyze *our concepts* of attitudes, not the attitudes themselves. Of course a proponent of the method may claim that that is *all there is* to the attitudes – after all, if there is *something* we know for sure about attitudes, it is that they are our way of talking. Perhaps attitude-talk is just that: talk, and there is nothing more to learn from attitudes than what our language reveals about them. But this may as well be a reason to *stop* philosophizing about attitudes completely. And this may be too quick. First of all, when we use attitude talk, we feel that it is about something more than just a manner of speaking. Especially when we talk about attitudes of human beings, we tend to believe that we are talking about *something* (why else are they so predictive?). And although we may wonder whether *other* people really have such things as beliefs, we feel pretty sure that *we* have them. Secondly, if we stop thinking about attitudes, we will not be in a position to explain why intentional talk is so pervasive in human life – nor why we are able to rely on it so much.

So how to go from here? An obvious strategy is to let the question be handled by science. If there is any institute that is supposed to figure out the correctness of our daily talk (idealized or not), to look at the "real nature" of things, it is science[14]. Science has told us that water is H_2O, that the sun does not really come up and go down, etc.. Sometimes science provides us with an explanation that seems counterintuitive at first, but often we correct – eventually – our intuitions and language in order to accommodate new findings.

Most philosophers indeed aim for some sort of vindication of attitude ascription through science. But what exactly is science going to vindicate? Should it find "real beliefs"[15] somewhere, most probably somewhere in the heads of people? Should it show that our belief-talk in fact corresponds to some lower level correlate, a certain kind of behavior perhaps, or some kind of mechanism? How much different can our scientific concept of 'belief' be from the ordinary concept? For instance, would we say that a behaviorist theory of the mental still

[14] As Sellars famously claimed, and has often been quoted in contexts like this one: "science is the measure of all things, of what is and what is not" (Sellars: 1963, p. 173)

[15] Or at least, some"thing" that has the same properties as beliefs; 'beliefs' in the everyday terminology. For instance, that we can derive logical conclusions from (sets of) beliefs, e.g. if I believe that supervisors should not forget to read all footnotes, and if I believe that Christian Illies is a supervisor, then I also believe that Christian Illies should read this footnote.

describes beliefs, or something else? If psychological behaviorism is correct, has it shown that beliefs don't really exist (because behavior is different from what we usually call beliefs), or has it shown 'the real nature' of beliefs? Or has science just not told us at all what *beliefs* are[16]?

Very roughly, there are two extreme answers to what the scientific attitudes are – and many positions in between them. On the one extreme we find the position that the attitudes don't really exist. This means that our talking about them is strictly speaking false – like talking about the sun rising in the morning. This is called eliminativism or anti-realism, and Paul and Patricia Churchland, and Stephen Stich are well known for defending some form of it (cf. Ramsey, Stich et al: 2001). On the other extreme we find the view that the attitudes do exist much like the way we talk about them: in that case, science has vindicate our attitude talk. Jerry Fodor is perhaps a good example of this position (Fodor: 1987).

Wrap-up

The three distinctions I have discussed are meant as pointers, rather than sharply defined concrete positions in philosophy. Most philosophers will hold some hybrid form of it. Dennett, as we shall see, lets all these distinctions collapse. And this, we shall see, has important implications, positive and negative, for his theory of function.

[16] The same problem is persistent in scientific accounts of phenomenological data (qualia). Science may tell us that pain is really nothing more than C-fibers firing, but this seems to fail to capture the real essence of pain: that it hurts. Similar problems arise with naturalistic accounts of morality (cf. Gerrans and Kennett: 2006).

3 The Essential Dennett

In this chapter I introduce Daniel Dennett's theory of mind, and its key constitutive elements: pragmatism and naturalism, united in his third person perspectivism. Integrated with the thesis of radical translation (Quine), I can set out Dennett's project and roughly place it in the attitude debate. On the basis of this, I derive the two standards that I evaluate Dennett's theory of functions with; standards, I argue, that every pragmatic naturalist should embrace.

Dennett's theory of mind counts as a controversial theory. It runs against a number of common sense ideas of what attitudes are, as well as against many philosophical theories that want to stay close to these common sense ideas.

For instance, when asked to locate a belief, many people will probably think about a brain, or a head of a person. Dennett denies this. Beliefs and desires are, in his view, in the eye of the beholder. Crudely put, any time it is useful for you[17], the beholder, or rather, ascriber, to treat some entity as having a certain set of beliefs, it has those beliefs, and it is in fact a believer.

Equally controversial is Dennett's thesis that we know ourselves in a way not principally different from others (we may know ourselves better than others, but this knowledge is not of a different kind). For Dennett, the whole mind can and should be understood from an outsiders perspective.

In philosophy, Dennett is rather controversial as well. First there is his style. Dennett's books are written in an accessible, pleasant, even cheerful style (something even his greatest "opponents" would be ready to admit), and as a result many of his books have found their way to a very broad audience all over the world. Books like *Darwin's Dangerous Idea* (1995b) about evolution and the mind, *Consciousness Explained* (1991a) about consciousness, *Freedom Evolves*

[17] Palmyre Oomen, in her response to the previous version of this dissertation, noted that false accusations are usually very useful to make, yet inherently *false*. As far as I know, this observation has not been discussed in the literature, but I take it that Dennett is only after 'epistemic' usefulness, i.e. the ability to predict and explain events, but we might want to add some principle of sincerity just in case.

(2003) and *Sweet Dreams* (2005) have been translated and sold all over the world. Dennett is also rhetorically strong – appealing to intuitions that work for him, and ridiculing those that work against him. (This is, by the way, a skill he shares with his greatest philosophical opponent, John Searle. Both are very good in addressing a broad audience in an accessible and rhetoric style. Philosophically they could not disagree more – their discussions are guaranteed to make an excellent read).

Dennett's strong rhetoric certainly does not only work for him. Many professional philosophers find Dennett's popular rhetorical style inappropriate and simplifying. It tends to make professional philosophers suspicious, and to ask him for more concrete details in the proper philosophical jargon, so that it can be tested and discussed with the proper (i.e. shared) conceptual tools. (The controversy of course already starts here – "new" (or non-mainstream) approaches often come with new concepts, or challenge the existing rules of discourse and the current paradigm; it is not always clear to what extent challengers of the paradigm may be forced, or free, to use different concepts and rules. This is especially problematic in a branch like philosophy, where it is not at all clear what the rules are, and to what extent they may be bended – the rules are part of the ongoing philosophical debate. Postmodernists and feminists can testify to this).

A second reason for controversy could be that the controversy is perhaps not so much about Dennett's position as such, but over some "stereotypical" version of it. Dennett's theory of attitudes is one of those theories that has gained the status of a "labeling position" in philosophy. Let me say a little bit more about "labeling" in philosophy. "Labeling" is an important argumentative strategy in analytic philosophy. 'Labeling positions' figure as idealized positions or ideal types that emphasize certain extreme characteristics of that position and enable a precise and systematic treatment of philosophical idea. In analytic philosophy, philosophy is often seen as an if-then science[18]: given certain premises or assumptions, certain theses must follow or cannot follow. A labeling position often figures as a kind of placeholder for the premises or assumptions. The philosopher defines the labeling position, usually in a idealized form, and specified in a number of precise theses. He then shows what follows from the position thus framed, or shows what does not follow. Theorists that seem to fall under some

[18] I learned the terminology from my former colleague Deniz Ogretir.

version of the label, can then see whether or not the reasoning is true for their position (i.e. if they support the specific theses, or not) – usually they will claim their position is a little different than the label, and can show in a rather precise way how they differ, and on what grounds the conclusion that follows from the thesis as embedded in the label, does not follow from their position.

Skepticism is an interesting example of a labeling position. The skeptic holds the negative position in philosophical debates: that there is nothing that we can know for sure, that everything is meaningless, that there are no ultimate values, etcetera. The challenge for non-skeptics lies in finding arguments that would persuade the skeptic to accept at least some kind of value or fact, some sort of secure starting point, that can then be used to build up a constructive theory (think about Descartes' *cogito*).

Dennett's instrumentalism is often treated as a skeptical position, a position that is to be challenged, rather than one that counts as a substantial position that may bear the final answers to questions about the mind. Now, on my interpretation, Dennett is not a skeptic at all. His theory is much too optimistic for that. But I do believe that his theory has gained a position in the philosophical debate that is similar to that of the skeptic. It may be the label or ideal type rather than Dennett's theory itself, that is most contentious, or at least plays the role of being a controversial position in the philosophy of mind. His theory counts as a position one should fight against, a position that threatens too many foundational assumptions.

Dennett has written two books about attitudes, *The Intentional Stance* (1987d) and *Brainstorms* (1978a). These books contain a number of papers that are central to his theory of attitudes. From the first there are: 'True Believers', 'Intentional Systems in Cognitive Ethology', and 'Evolution, Error and Intentionality' (resp. 1987f, 1987b, 1987a). From *Brainstorms*, there is the classic paper 'Intentional Systems' (1978c), and 'Conditions of Personhood' (1978b). Also, the journal papers 'Real Patterns' (1999) and 'The Interpretation of Texts, People, and other Artifacts' (1990) are important for understanding Dennett's theory of attitudes, his Intentional Systems Theory.

Contrary to his work on consciousness and evolution, Dennett's theory about attitudes has not gained a lot of popular interest (except, perhaps, via his *Kinds of Minds* (1996)), probably because it is a typical philosophers' topic. Or perhaps because Dennett himself has moved on to the 'hard' problems of consciousness, freedom, evolution, and culture. This is not to say that his theory of attitudes, Intentional Systems Theory, is outdated or no longer philosophically interesting. On the contrary, Intentional Systems Theory is the pillar of his broader theory of the mind. Moreover, it is a continuous source of inspiration for many cognitive scientists, both in AI and in psychology[19] and it is still discussed frequently in the philosophical literature.

Note the contrast between the words "inspiration" and "discussed frequently" in the last sentence. Indeed, for many scientists, especially those working in cognitive science, Dennett is not controversial at all. We will see some of that influence on and from cognitive science in chapters 6 and 7. It is Dennett's *naturalism* that partly explains why cognitive scientists find him interesting, and why philosophers, at least non-naturalists, are critical towards his position.

Naturalism is usually defined as the philosophical position that says that acceptable methods of justification and explanation are continuous, in some sense, with those in science. (The Cambridge Dictionary of Philosophy: 2001, 596). We may read this in a weak form. Then it means that philosophical results cannot plainly contradict scientific results, but it allows for the idea that there may be phenomena that are forever beyond scientific understanding. Most philosophers endorse at least a weak form of naturalism - but I prefer not to call this naturalism. The hot issue, especially in the philosophy of mind, is whether there are facts or phenomena that cannot - in principle - be grasped by scientific means or, more generally, from a third person perspective. Dennett believes that the mind can be fully understood from the third person perspective, most notably science, and that makes him a naturalist in my terms.

3.1. Dennett's third person view on the mind

Dennett naturalizes the mind by claiming that the whole mind can be described and explained from the third person perspective. For Dennett, when our third person perspective sources about a possible mind-haver are exhausted, there is nothing more to say about a mind. Let's call this the exclusiveness thesis: there is nothing more to say about the mind, than what knowledge from the third person perspective could possibly tell us about it. The third person perspective,

[19] See Ross, Brook et al: 2000 for an overview of research inspired by Dennett's philosophy of attitudes.

according to Dennett, is our best and only reliable access to a mind. Dennett denies that we have privileged access to our own beliefs (Dennett: 1991a, 96), and claims that introspection is not a good way to know what goes on 'in our heads'. For instance, our 'first person perspective' (or, how things seem for us) wrongly suggests that vision is like seeing pictures in the head (for an argumentation why this would be wrong, see Dennett: 1991a, 55 and before). Reports of how things seem to be to us are important data for science to account for and to explain, but must never be taken to be authoritative[20]. The objective third person perspective of science performs better – we do make mistakes from the third person perspective, but these are traceable and corrigible, like every hypothesis in science.

According to Dennett, we are also not introspectively authoritative about the real content of our thoughts. Just think of those cases in which you thought you believed or liked something, but turned out to be wrong about your own beliefs after giving it some further thought or after being challenged by someone else. For Dennett, you are an interpreter of your own behavior, very similar to the way you interpret behavior of other people (Dennett: 1991a, 246). This self-interpretation is mediated by (public) language and based on the same kinds of facts as the ones we use when we interpret other people (I will say a lot more about interpreting behavior in the next section). You construct a theory of what you believe and desire, and who you are, on the basis of your positive and negative responses to your environment. (The only difference is, of course, that you have a lot more data available than other people – in this sense, you have some privilege).

So, many facts that seem first person perspectivist at first, turn out to be third person perspectivist on Dennett's account. For Dennett, the first personal phenomena that seem to be left unexplained (for instance, the "qualitative feel", as they say, or "qualia"), have little relevance to our understanding of the mind and can thus be ignored (Dennett: 1993).

[20] Dennett argues that first person perspective data can be very well studied, by letting subjects report their experiences. Dennett calls this heterophenomenology. We can ask a person what he (thinks he) feels and thinks, as precisely as possible, and use these reports as data for further research. As "texts" they are subjective data that are third-person perspectivized. First person perspective data is not taken to be as an authoritative, autonomous source, but as data to be explained (Dennett: 1991a).

Many philosophers restrict themselves to either the theory of attitudes, or the theory of consciousness. Dennett, however, wants to show the connections between the two and aims to construct a theory about both from the same third person perspectivist principles. For Dennett, the theory of consciousness *follows from*, and is based on, the theory of attitudes (Dennett: 1998, viz. 355 and Dennett: 1969). Loosely put: having consciousness is a complex way of having attitudes. Some creatures can be ascribed attitudes, but not consciousness. (compare, again, Dennett's concept of personhood: a person is a creature that can be ascribed intentional attitudes *and* the capacity to do something with those attitudes: reflect upon them, and communicate them). Beliefs, then, are not necessarily conscious beliefs – beliefs can be had by creatures that have no consciousness at all.)

For Dennett, it is important that we see the mind as an evolved capacity of organic beings. Somewhere in the process of evolution, organisms started to show *intentionality*, but not yet consciousness. Creatures that show intentionality are creatures that can be ascribed beliefs and desires, creatures that in some way represent their environment, and have certain preferences. Even very simple organisms, like plants and even amoebae, have intentionality in this sense - though in a primitive way. Dennett believes that (human) consciousness, in the end, evolved from these very simple beliefs and desires. Let me quote Dennett at length:

> There was a time, before life on earth, when there was neither intentionality, nor consciousness, but eventually (...) simple organisms emerged. (...) Were they conscious? Did their states exhibit intentionality? (...) One family of intuitions [Dennett's family] is comfortable declaring that while these earliest ancestors were unconscious automata, not metaphysically different from thermostats or simple robotic toys, some of their states were nevertheless semantically evaluable. These organisms were, in my terms, rudimentary intentional systems, and somewhere in the intervening ascent of complexity, a special subset of intentional systems has emerged: the subset of conscious beings. According to this vision, the intentionality of our unconscious ancestors was as real as intentionality ever gets; it was just rudimentary. It is on this foundation of unconscious intentionality that the higher-order complexities developed that have culminated in what we call consciousness. (1998, 358)

So we have creatures that have intentional states (beliefs, desires, etc.), and we have creatures that have *conscious* intentional states. Conscious intentional states, then, are what is called "higher order intentional states": beliefs about beliefs, beliefs about desires etc. (1991a, 309). When you consciously entertain the belief "This lecture is boring", you are, technically, having an unconscious belief about a belief: "I believe that I believe that the lecture is boring". Consciousness is, in other words, higher-order thinking, or *higher order intentionality*[21]. Precisifying beliefs in this way is important to explain behavior (which can be "caused" by conscious as well as unconscious beliefs), to understand what consciousness exactly is, and to explain how it could have evolved from simple intentionality.[22] This thesis needs only Dennett's theory of attitudes, not his more complex (and more controversial) theory of consciousness. Because attitudes are the simple case, I will not further discuss his views on consciousness.

By taking an evolutionary approach, Dennett hopes to shake off deep philosophical problems about the mind, most notably, the mind-body problem or the question how physical things can (also) be intentional and conscious. Minds have evolved from very simple, almost mindless entities, to the kinds of minds human beings have.

3.2. Interpretationism and radical translation

Dennett's evolutionary approach tells us something about how minds came to exist, but what can we say about the *contents* of mental states? The mind-haver is not authoritative about the content of his own mental states. So who is? How can we tell what someone or something believes? How can we tell if someone has beliefs at all? Dennett uses the interpretationist framework to answer these questions.

Child defines interpretationism as follows:

[21] On a very general level, Dennett's views can be shared amongst the HOT-theorists of consciousness. HOT stands for Higher Order Intentionality.

[22] More about the relation between attitudes (content) and consciousness in Dennett's Ph.D. Thesis (1969), and in Dennett: 1998.

> Interpretation is the process of ascribing attitudes to an individual on the basis of what she says and does. When we interpret someone, we aim to make sense of her by attributing beliefs, desires, intentions, emotions, and other propositional attitudes – attitudes in the light of which her behavior is intelligible as, more or less, rational action. (Child: 1994, 7)

Interpretationism is heavily inspired by Quine, especially his thesis of radical translation. What is this thesis?

Consider a simplified version of a famous thought experiment from Quine (cf. 1960, pp. 26 and further). Imagine you are a traveler to a distant land, and encounter a tribe that speaks a language that you cannot understand. Let's also assume that the members of the tribe are friendly and cooperative. You, of course, want to learn everything about them, and, therefore, have to learn what they mean when they use their strange words. What will you do? You might start by taking a stick from the ground and say "stick". They may answer by clapping their hands and saying "kitchu!". You hypothesize that 'kitchu' means 'stick' (and that clapping their hands means that they understand what you mean). Then you point at a tree, and say "tree". The tribe members respond by saying, again, "kitchu". Hm, you wonder, perhaps 'kitchu' does not mean 'stick', perhaps it means 'wood'. You point at, say, your wooden shoes (the previous destination for your travels was Holland) and say, questioningly, "kitchu?" The members of the tribe fold their hands and agree: "kitchu!" Ah, apparently the second hypothesis was right: 'kitchu' means 'wood', and folding the hands means 'yes!' (or perhaps not, we may want to find out if 'kitchu' does not mean 'wood that some strange traveler points at...')

You go on for a while, and learn the words they attach to certain objects and events. Or in Quine's terms: you are after the *stimulus meaning* of their words and sentences. 'Kitchu' and 'wood', in Quine's terminology, turn out to have the same stimulus meaning (Quine: 1960, 33). It is called stimulus meaning, because you determine the meaning of a word or sentence on the basis of the stimuli that are assented to when uttering a certain word or sentence (in the case of "kitchu!": the stick, the tree, the wooden shoes, in relation with the folding hands). The process in which you try to find out the stimulus meaning of the words and sentences of the tribal language is called 'radical translation'.

Note that we acquire the stimulus meaning entirely from the third person perspective – there is no other way to understand the meaning of the words in the tribes language.

> "In psychology one may or may not be a behaviorist but in linguistics one has no choice. Each of us learns his language by observing other people's verbal behavior and having his own faltering verbal behavior observed and reinforced or corrected by others. We depend strictly on overt behavior in observable situations. (...) There is nothing in linguistic meaning beyond what is to be gleaned from overt behavior in observable circumstances" (Quine: 1987, 5)

According to Quine, the example of radical translation reveals what meaning fundamentally is - and what it is not. Philosophers like Dennett and Davidson have followed him in this, although they call the process of finding out the meaning of words and sentences (or other meaningful items, like objects, or signs and gestures) *interpretation* and even though they have not adopted his strict behaviorism.

I have simplified Quine's thesis of radical translation and his theory of meaning for reasons of clarity[23]. It should help to get some basic points clear.

First, at least according to Quine and Dennett, there may be an end-point to the process of translation or interpretation. If so, this is also the end of determining what the stimulus meaning is – hence, what is meant. In Quine's example, the radical translator finds out that the natives use the word 'gavagai' when they see or point at a rabbit. But 'rabbit' and 'gavagai' may have exactly the same stimulus meaning but a different reference. Perhaps they use 'gavagai' to refer to 'a stage of rabbit' (a kind of time-slice of a rabbit), or to 'integral parts of rabbits' or the universal 'rabbithood' (1960, 52). It is quite unclear how we could find out *what* the correct answer is. According to the interpretationist, in such cases there may be none (here we have, again, the implications of their 'pragmatism'), not even for the natives themselves. This is called 'the indeterminacy of radical translation', or, for Davidson and Dennett, 'the indeterminacy of radical interpretation' (Davidson: 1984, Dennett: 1987c, Dennett: 1999)[24].

John Searle has argued that this thesis of underdetermination only works if one presupposes that psychological meaning does not exist – and considers that

[23] Quine's thoughts on translation and indeterminacy are laid out in his 1960, 1970, and 1987.

[24] Davidson is more optimistic than Dennett about the number of cases in which such indeterminacy may arise. The reason lies in Dennett's more liberal view on what belief is. Davidson only counts conceptual beliefs (roughly: conscious beliefs) as real beliefs and conceptualized beliefs are more specific (Davidson: 2001).

a reduction ad absurdum (Searle: 1987). The stimulus meaning of two sentences may be the same, but the psychological meaning different (*I* know whether I mean rabbit, or rabbit stage when uttering *gavagai*). Meaning, Searle contends, is only underdetermined from the third person point of view, not from the first person point of view. But it is exactly the point of the interpretationist that there is no first person point of view – as Searle wants to have it. For the interpretationist there is nothing more to meaning than interpretation of stimulus meaning[25]. So Quine's thesis is not restricted to linguistic meaning. It also holds for *psychological* meaning (if it even exists). If you want to find out what someone (or something) means, you have to start interpreting: our best and *only* way to get to meaning. If only because we ourselves have learned the language, and the meaning of words and sentences, in a process similar to the radical translator. Just think about your mother pointing at things, giving names to it, and you trying to use the words you learned in different contexts (Quine: 1987)[26].

It remains an issue of debate whether indeterminacy is a drawback of interpretationism. Dennett thinks it is not. He thinks that the idea of precise and determinate content is wishful thinking. He claims that it is exactly our ability to use language that may have created the *illusion* that content is determinate (the "illusion of specificity", see Dennett: 1987b, 255).

The second (not unique) characteristic of the thesis of radical translation is that the meaning of things is determined *in relation to* the meaning of other things. We found out what 'kitchu' means by testing it in different contexts (pointing at a stick, at a tree, at wooden shoos), and by relating it to other terms (the folding of the hands that supposedly means 'yes we understand'). This is called *meaning holism*: the idea that meaning is always determined 'holistically', that is to say: in relation to other meanings.

The third consequence of the thesis of radical translation is that our ascriptions ('translations') are regulated by the principle of charity. That is to say, when we make a translation or ascription of attitudes, we must assume that the subject is rational (bound at least to the basic rules of logic) and interpret his utterances

[25] Quine is agnostic with respect to psychological meaning. But according to Searle, and this seems right, Quine's thesis only works on the assumption that there is no psychological meaning.

[26] According to Quine, the *whole* language can be understood by means of this process of translation, not only so-called observation sentences, but also logical particles and analytic sentences (Quine: 1987).

in a way such that attributions do not contradict each other and form a coherent whole (Quine: 1960[27], Thagard and Nisbett: 1983). I will say more about the crucial role that the principle of charity plays in Dennett's theory of the stances in 3.4.

3.3. Interpretationism and science

Since the dawn, and even the decline, of psychological behaviorism, mental states such as believing, desiring etc. are considered suspicious parts of the scientific view of the mind. Mental states, behaviorists thought, are internal, *introspective* facts, and as such will forever elude science (cf. Watson: 1919, Skinner: 1938). The strict behaviorist program effectively put an end to the method of introspection in science, and even today mental state terminology is treated with extra care in the psychological sciences. Yet modern psychology is much more liberal than the original behaviorists thought justified (see Allen and Bekoff: 1997 for a brief but excellent overview of the history of behaviorism, see also 4.1).

Following psychological behaviorism, Dennett's teachers Ryle and Quine constructed the philosophical counterpart of behaviorism: logical or philosophical behaviorism. Logical behaviorism says that mental states, and the ascription thereof, are part of everyday language, a piece of "dramatic idiom" that can never be part of science. Gilbert Ryle (1949) famously split between causal and "conceptual" explanations that we can give to mental phenomena (see also 4.1). These two types of explanations have to be rigorously separated: if we give a causal answer to a conceptual question (or vice versa), we make a category mistake.

Quine follows a similar route. The thesis of radical translation was meant primarily to say something about certain *kinds of sentences*. Quine (1960) emphasized that intentionalistic language is "intensional" whereas scientific (physical) language is "extensional". Extensional sentences refer directly to some state or event in the world. The terms in the sentence *extend*, as it were, to the world. Sentences like "Peter believes that *x*", or "Suzy hopes that y", are, by contrast,

[27] The original formulation of the principle of charity is from Wilson, construing the principle in terms of truth rather than rationality: "We select as designatum that individual which will make the largest possible number of ... statements true" (Wilson, quoted in Quine: 1960, 59)

intensional[28]. The meaning of a scientific statement can be specified by reference to "the world", whereas the meaning of a psychological statement (S believes that p) is determined holistically by means of (radical) translation. Intensional sentences, Quine thinks, can therefore never be a proper part of the scientific vocabulary.

Dennett firmly rejects *psychological* behaviorism à la Skinner because a Skinnerian psychology would not be a psychology *at all* (1978d). His relation with *logical* behaviorism is, however, a bit more complex.

Dennett agrees with Ryle and Quine that attitude-talk is of a different kind than 'proper' scientific talk. He sees the importance of making a distinction between the extensional and the intensional ('meaning'), and adopts Quine's meaning holism. Also, Dennett agrees with Quine that we cannot make straightforward *science* of statements about intentional states and the like. And Dennett also thinks that introspection is not a valid method for science: we, like the radical translator, will have to work with data from the third person perspective, behavioral data, and interpreted as actions (behavior "under a certain description", cf. Davidson: 1980). In this general sense Dennett is, like Quine, a logical behaviorist.

But he does not accept the conclusion Quine and Ryle draw about mental states: that they are *just* talk. He believes the predictive force of attitude talk should not be put away that quickly, and that it can be put to good use for science, as a method for psychology and other "hermeneutic" sciences (especially economics, and evolutionary biology). The practical utility of the intentional stance should not be put away so quickly. In chapter 4 I shall deal in detail with the 'science of the intentional stance'.

3.4. Characterizing intentional explanations

In his methodological writings, the sociologist Max Weber proposes a dual methodology for the social sciences. Weber thought that on the one hand, the social scientist, like any scientist, has to look for causal laws – defined by him as "the probability that an event will be followed or accompanied by another event"

[28] There are more intensional sentences than those that refer to intentional states, amongst which modal sentences and quantifying sentences. For Dennett's project, however, those that refer to intentional states are most relevant.

(see Ritzer: 1996, 224). On the other hand, the social scientist has the advantage over the natural scientist that he can also use the hermeneutic (i.e. interpretative) method to *make sense of* and *understand* the subjects of these causal patterns, i.e. to grasp their perspectival states (the 'hermeneutic method', after Dilthey). In other words, for Weber, causal laws explain behavior (*erklären*), whereas hermeneutics is used for understanding actions (*verstehen*).

Many social philosophers who work in the continental tradition, have accepted the hermeneutic method as a scientific method of sorts, indispensable for any science that deals with meaningful action[29]. Curiously enough, this idea has hardly gained any ground in *analytic* philosophy of action. Action, in this tradition, is seen as a phenomenon that has to be understood in causal terms– which attaches it directly to the mind-body problem and other persistent philosophical caveats (see also chapter 4).

Intentional Systems Theory, as defended by Daniel Dennett, seems to be an interesting exception. Dennett uses interpretationism as a way to say something about the content of attitudes from the perspective of the agent (e.g. what he can know, what he believes, what he may desire), in the sense of the weak first person perspective (see 2.1), without being committed to introspectionism.

There is some debate about whether Dennett takes intentional explanations to be causal explanations (see Elton: 2003 and chapter 4). I take Dennett to claim that action-explanations are *not* causal explanations. The causal work, whatever that may be, is done at some physical or bio-neurological level. But if we are explaining or predicting someone's behavior in terms of *reasons*, we are trying to *interpret* his actions. We are engaged then in a hermeneutic enterprise, not a causal explanatory one, nor a statistical one. What is important to emphasize is that taking an interpretative stance (the intentional stance and the design stance as we will see), is explaining and predicting *the very same* event from a *different* explanatory framework. We may thus explain behavior from the physical-scientific framework, or describe it under an action description from a hermeneutic-interpretative framework. I will explain this in 5.1.

[29] Cf. Ritzer: 1996: 222-225. Note that this claim necessarily goes with an extension of the definition of science. By 'science' we usually refer only to the 'real', physical sciences. The hermeneutic method is not a scientific method in this traditional or physical sense. The claim of the interpretivist is that interpretation is a different but valid method in science (hence redefining what science is), even though it is not 'scientific' in the traditional sense.

What, then, are the typical characteristics of intentional explanations? When we want to know what someone or something believes or desires, we interpret his behavior, like we translated the utterances of Quine's natives in the far away land. Our best possible interpretation of behavior of an agent will give us what the agent believes and desires. Dennett calls the process of interpretation of behavior in terms of attitudes 'taking the intentional stance'. We look at the behavior of the creature from the intentional point of view and we view the creature as a creature that has attitudes. When we take the intentional stance, we ascribe certain beliefs, goals and desires to the system, which we derive from the system's epistemic needs, its relation to the environment, behavioral descriptions, its perceptual capacities and so on. But this can be done only if we *assume* that the system is *rational*: the *principle of charity*. We have to presuppose that it follows certain rational norms. In §2.1, I already mentioned that attitude-explanations are subject to rational evaluation. Some philosophers think that the rationality lies in the entity that *has* the attitudes (see below). An interpretationist thinks that the rationality instead comes from the activity of interpreting: the interpreter *assumes* that the entity is rational, and his hypotheses of what the entity believes and desires is based on this assumption. The interpreter, in other words, is *charitable* to the interpreted agent and the assumption of rationality constrains the interpretation. So, according to the interpretationist, the interpreter *rationalizes* actions. If we cannot make rational sense of the behavior of the entity at stake (if it is utterly irrational), we have to abandon the intentional interpretation. Note that we do not have to assume perfect rationality, people make mistakes every once in a while, and are perfectly understandable nevertheless. Minimal or bounded rationality is usually the preferred term in these contexts. Nevertheless, we have to assume that intentional agents follow the rules of rationality most of the time (Thagard and Nisbett: 1983, Cherniak: 1986). In this sense, an intentional explanation is necessarily a rationalizing interpretation. It is through rational interpretation, and *through interpretation alone*, that we can determine what someone believes, and whether it is an intentional being, an agent. If we can't make sense of an agent anymore, we have to abandon intentional interpretation. It that case it is not an intentional system anymore.

This ideas is captured in Dennett's famous thesis that:

> "*[A]ll there is* to being a true believer is being a system whose behavior is reliably predictable via the intentional strategy, and hence *all there is* to really and truly believing that *p* (...) is being an intentional system for which *p* occurs as a belief in the best (most predictive) interpretation" (1987f, 29)

Intentional states, then, are *ascribed states.* If I successfully interpret your behavior on the basis of ascribing certain attitudes to you, you *have* those attitudes. Another way of saying this is that attitudes are *constituted* by an act of interpretation, rather than *discovered.* In the terminology of chapter two, *attitudes themselves* are constituted by *attitude ascription.* To the interpretationist, attitudes are thus *mind-dependent.* They exist only from the perspective of the interpreter and are, in a sense, *created* by him. Many philosophers conclude that Dennett must be an anti-realist about attitudes: attitudes do not really exist (for a short characterization of Dennett's 'ontology, see 3.6).

But as every historian knows, the epistemic status of an interpretation is questionable: it is an *interpretation.* How do we tell good interpretations from bad ones? How can we tell whether we have interpreted the behavior correctly - how do we know if we attributed the *right* beliefs and desires to the creature? Isn't interpretation an *approximation* of what someone believes and desires?

A first answer is that a good interpretation makes good sense of the behavior at stake, and does not contradict the behavioral evidence we have. The interpreter, like the historian, has to make a 'case' for his interpretation. Interpretations have to be compatible with the external facts (the environment), should be internally consistent, and so on.

But once we have exhausted our sources, when there is no more behavioral data available, there is nothing more to find out. Compare, again, Quine's field linguist who tried to find out what the natives mean when they use certain words. At a certain point the translation comes to an end, and at that point there is nothing more to say what someone means, believes or desires (cf. Dennett: 1987f, 28-29). This means that psychological meaning, 'content', can be indeterminate; just like linguistic meaning is (Dennett: 2000).

Still, even if we grant that there are good interpretations and bad interpretations, we may still wonder whether interpretation is a good method at all. As we saw, we have to make a risky assumption, namely that the system at stake is rational. And what we basically do is find a good story that fits the behavioral data, and this involves a lot of guesswork. And, as we saw, sometimes there are more good stories to be told. This can hardly be called a good method for science – why would Dennett, as a naturalist, defend such a method?

Recall that attitude explanations and attitude predictions work remarkably well in daily life – even though they are certainly not fail-safe. If I make a promise with you to pick you up at the airport, and if you know that I am a trustworthy person, and that I want to keep my promise, you can reasonably *expect* me to pick you up. This is a reasonable prediction. But it is no *guarantee* that I will be there – something else may turn up, or I may get a flat tire. This is typical for attitude-explanations: that they work remarkably well, but only under ceteris paribus conditions. So we cannot expect intentional predictions to have it always right. A good intentional prediction is not a prediction that precisely predicts what will happen. The value of an intentional prediction rather lies in its efficiency and usefulness.

Compare intentional explanations with physical explanations. From a physical point of view, we can theoretically predict what will happen with great certainty and accuracy. But just imagine how much physics and neurophysiology it would require in order for you to predict that I will be at the airport. You will not be at the airport in time if you try to do that! From the intentional stance, you would have predicted that I would be at the airport in a flash. It comes almost *natural* to you. Quite literally, even, as we will see: the capacity to take the intentional stance is deeply embedded in our mind. And this should not surprise us: the intentional stance is a life-saver.

Dennett's justification of the use of stances on the basis of their relative usefulness is sometimes called *instrumentalism*. Instrumentalism basically says that the theory that helps you explain things in the most useful way, is the *right* theory. The value and justification of the intentional stance then depends on the purposes of the interpreter, and the task at hand. Roughly, if you want a quick prediction, you take the intentional stance, and if you want a precise prediction, you take the physical stance. If you want to survive in the jungle, use the intentional stance. If you want to make dynamite, or build an airplane: use the physical stance (or the design stance).

Dennett believes that science can and should profit of the strength of the intentional stance (see chapter 4 for the special characteristics of the intentional stance). What matters is that Dennett believes that the intentional stance can be useful for science, as a method, and that its legitimacy would lie in an instrumentalistic (pragmatic) justification of it.

3.5. Pragmatic foundations

We have seen that Dennett's theory of attitudes is strongly influenced by naturalistic considerations, but its instrumentalism gives it a significant pragmatic flavor too. In fact, I see Dennett as defending ultimately a pragmatic theory of the mind.

In my view, Dennett founds his naturalism on pragmatism - and it is perhaps exactly the combination that makes his position so controversial. What is pragmatism? Rorty's *Consequences of Pragmatism* gives a slightly biased, but good overview of pragmatic positions, and of the strict conclusions that he thinks should be drawn from it (see Rorty: 2001). Rorty describes pragmatism as the position according to which the search for Real Truth, Rationality, and Goodness should be given up, to be replaced by questions (and answers) that help us to cope with the world.

He counts as pragmatist philosophers not only Dewey, Pierce and James, but also philosophers like Davidson, Quine, Sellars, Wittgenstein, and Dennett. And indeed, these philosophers seem to be motivated by a certain tiredness with the "deep" metaphysical" issues, that they deem quite meaningless. And indeed, Dennett writes:

> I am shy about drawing ultimate conclusions about Reality, Truth, Meaning, Time, Causation, and the other grand topics of metaphysics and epistemology. (...) I take myself rather to be just working out some of the more surprising implications of the standard scientific picture. (...) Like most cognitive scientists, I'm prepared to take my chances with conservative, standard scientific ontology and epistemology. (...) My "scientism" comes in the form of a package deal: you think you can have your everyday science and reject my "behaviorism" as too radical? Think again. (Dennett: 1995a, 204-205)

Many pragmatists share a wish to retreat from 'deep' metaphysical issues to language (and the claim that there is nothing deep 'behind' language). Contrary to Wittgenstein, Davidson and even Quine, Dennett is not so much focused on language[30], but rather on reformulating long-lasting philosophical questions

[30] But Dennett's approach has some clear verificationist roots. Verificationism is a position in the philosophy of language, and says that only that what is experientially verifiable is meaningful, e.g. the sentence "Everything has just doubled in size (including measuring sticks)" is not

(especially about the mind) into questions that are (intersubjectively) answerable. Dennett is a kind of no non-sense philosopher – evading, in some sense, these deep metaphysical issues, and constructing theories and questions that one can work with. If a certain question is at heart unanswerable (and we may of course disagree over when this is the case), it should be rephrased into one we can meaningfully answer. For Dennett, conceptualization is operationalization, and a philosophical question is always -partly- a question of method. Often, but not always (as in the case of the conceptualization of 'personhood' as I discuss below) his approach is to rephrase philosophical questions into questions that could be answered by science, or in a way that remains in the spirit of the scientific method (cf. Dennett: 1996 for a clear example of this approach).

It is in this sense that his pragmatism and his naturalism come together: science counts as the most “practical” way to deal intersubjectively with problems. (it is also in this sense that Dennett has acquired somewhat the position of the skeptic in the philosophical debate). According to this interpretation Dennett's pragmatism underlies and justifies his naturalism.

3.5.1. Concepts as tools

Let me give two examples that give an indication of Dennett's pragmatism and naturalism. At some point Dennett asks how we could possibly answer the question what a vulture thinks and feels when he smells and subsequently eats his dead prey? This question seems unanswerable: for there is no way we could get inside the vulture’s head, and know how it experiences the world. Many philosophers indeed claim that because we cannot gain access to the vulture’s experience, there are facts about it (experiential facts) that will forever elude us (cf. Nagel: 1974). The typical Dennett move on these kinds of issues is to start with what we *can* know about the vulture’s experience. For instance, from the fact that vultures live on dead carcasses, we may safely conclude that the vulture at least has no terribly negative experience when he eats it. The smell of death probably even increases the vultures' appetite.

meaningful because there is no way by means of which we could experientially verify the truth of such a statement. Cf. Lycan: 2004. Verificationism is not popular anymore, but some of the intuitions behind it are still alive in philosophy, including Dennett’s philosophy.

Now, this simple example is of course fairly trivial. But it does illustrate the constructive and pragmatic way of approaching these issues. And the example can be easily extended. In his reply to Nagel's claim that no third-person knowledge could tell us what it was like to be a bat (Dennett: 1991a, 441-448), Dennett argues that we can say quite a lot about the experience of a bat, for instance on the basis of its physical constitution (a colorblind animal cannot experience the color orange) and inferences about its needs (food will usually be experienced positively, and predators negatively, to give some obvious examples).

A second example. What is a person? Many philosophers define a person as a creature that has a first person perspective. Recall from chapter 2 that it is not very clear what it means to have a first person perspective, but in this case it means something like having a self, or being able to conceptualize yourself as yourself. (cf. Baker: 2000, Baker: 1998, Shoemaker: 1996). This claim usually goes with the thesis that the final answer of personhood lies in the person itself: only the one having the first person perspective truly knows whether he or she is a person.

Dennett's approach is different. He starts by pointing out that there is a practical need to construct a concept of a person, even though it seems hardly definable in terms of necessary and sufficient conditions (Dennett: 1978b, 268). We need the concept to answer certain questions, mainly moral in nature or very relevant for moral questions. Questions like: Is a two year old child a person? A five year old? Could there be non-human beings that are persons? Is a chimpanzee, for instance, a person? Can we hold every person accountable for his or her actions, or is there more to it? At what point does someone have so little control over him or herself that we cannot consider him to be a person (i.e. responsible for his actions) anymore?

For Dennett, a conceptualization of 'person' that is unable to answer such questions is quite useless, not to say, wrong, empty and meaningless. Why would we even bother defining personhood, if we cannot operationalize it? Let us assume for the moment that persons, indeed, have some special relation to themselves – be that a first person perspective or something else, self-consciousness perhaps. Something like that seems plausible, given that we want to attribute responsibility to persons, and responsibility at least requires some kind of self-reflection. How can we tell if a certain organism, or machine, has it? These are for Dennett the kinds of questions that we should be able to answer with a notion of 'person'.

Dennett thinks we need a conceptualization of the first person perspective from the third person perspective – a conceptualization, in other words, that everybody could use to determine whether someone else (or even oneself) is a person. Amongst these external criteria are whether someone shows the capacity to think about his own beliefs (has higher order intentionality), whether he is able to deceive (a capacity that is quite rare amongst natural organisms) and whether he is able to reflect on his will (which is equally rare) (Dennett: 1978b.). And these criteria are not checked on the basis of internal facts (my 'knowing' that I reflect upon my own will, or my 'knowing' that I can deceive), but on the basis of facts that appear from the third person perspective[31]. What matters is Dennett's claim that personhood is something that must and can be perfectly well understood from the third person perspective - as long as we settle for a pragmatic concept of personhood, which is, according to Dennett, the only conceptualization one should aim for. Or, more positively, we may say that the question of personhood is too important to be answered by merely a theoretical definition.

The personhood example is interesting because it shows the primacy of pragmatism over naturalism. According to Dennett, there are no hard scientific facts that could determine whether someone has a first person perspective in the sense that there are no brain-facts or neurological facts that we could use to determine whether something has a first person perspective – is a person[32]. 'Person', for Dennett, is an everyday term, not a scientific one. But, and as such,

[31] One may now think that Dennett may have given a few reasons why we can know 'other minds', or at least make educated guesses about other minds, but that, in the end, only the person at stake is able to say what he thinks, feels, and is able to do. Some claim that the question whether someone is a person can only be answered by the person himself and the experience of the vulture is only known by the vulture itself. Or it is, perhaps, what they call, "a metaphysical fact". Dennett disagrees. Perhaps there are internal, or 'metaphysical', facts that would give a definite answer to these questions, but it is hard to see how these could ever play a role in our decision to hold someone responsible for a certain action. For Dennett, these facts are no facts at all. This approach is very similar to that of Strawson when he conceptualizes responsibility in terms of reactive attitudes. For Strawson (1968), the justification of holding someone responsible lies in our having the proper attitude (blame, praise, resentment etc.) towards him. Responsibility and personhood, in this view, do not depend on an internal fact (metaphysical freedom for instance), but can be determined from an outside perspective.

[32] We may, of course, expect a certain complexity in brains of persons but there is no brain area, so to say, that glows up red whenever someone uses his first person perspective.

'personhood' is an important concept for us (1978b, 268). If only because it plays a crucial role in morality: if we characterize a creature as a person, we immediately give him a place in the moral domain, as a creature that can be held responsible. So, despite the fact that we cannot give a scientific conceptualization of 'personhood', the term is not abandoned or eliminated.

The personhood example also shows that useful (i.e. practically relevant) answers are usually answers that we can fully answer from the third person perspective - paradigmatically but not necessarily science. If we have to answer the question, for instance, whether some creature is to be considered a person and draw practical consequences from it (give him certain moral or legal rights for instance), we (should) demand that this choice can be justified as objectively as possible, which means at least shared intersubjectively and thus third personal. Science, by most considered to be an objective arbiter with respect to facts, is weak when it comes to norms and values. It is for this reason, I submit, that Dennett chooses a non-scientific conceptualization of personhood. Dennett's approach to concepts as tools is paradigmatic for his pragmatic approach towards philosophical issues.

Dennett is without any doubt a naturalist. His way of approaching questions of mind, his attempt to justify intentional explanations within science, his third person perspectivism are all clear indications of that. According to my interpretation, however, Dennett's *ultimate* justification of naturalism lies in pragmatism and the third person perspectivism that underlies it. Third person perspectivism thus can be seen as a realization of pragmatism, and naturalism is a favorite, but not the only, form of third person perspectivism. Furthermore, his pragmatism is not necessarily realized by naturalism. Take the case of the conceptualization of personhood, which is clearly third person perspectivist, but can hardly be called full-blown naturalistic (it is continuous with science, but does not propose a scientific method for discovering what personhood is). This means that the ultimate justification for a theory of ascription lies in pragmatic considerations, not in scientific validation per se.

3.6. Interpretationism and some ontological issues

At this point I think it is useful to contrast interpretationism with a main rival: intentional realism. Intentional realism and interpretationism mainly fight over ontological issues. Dennett's interpretationism is usually seen as an anti-realistic theory of attitudes. The reason for this, of course, is that attitudes only exist from the point of view of the observer, the interpreter. Intentional realists, contrary to interpretationists, think that attitudes *do* exist, independently from our ascriptions, and that our ascriptions of attitudes to entities in a sense correspond to something in the world, usually but not necessarily: in the heads or brains of people.

There are many ways to be an intentional realist. Some are conceptual, others are naturalistic, and some are both. For instance, Lynne Rudder Baker defends a form of conceptual intentional realism (Baker: 1995). Her intentional realism is interesting in relation to Dennett's interpretationism because she, like Dennett, does not think that attitudes are in the head. But, contrary to Dennett, she thinks we can take everyday language more or less at face value: our daily explanations of actions *count* for something, and are legitimate whether or not science finds a place for them. Baker's theory is called practical realism, it is mainly a metaphysical theory (that is, it is about what exists) and it has everyday language as its starting point.

But in this paragraph I want to discuss a different form of intentional realism, that of Jerry Fodor, because he makes the clearest contrast. Very roughly, Fodor thinks that our everyday use of attitude-talk will be vindicated by science, that is, he believes that our daily talk is more or less correct. Attitude ascription will turn out to be *about* attitudes themselves. So, when I correctly ascribe to you the belief 'that you think you feel like going to a party on Friday', this means that there is some item, in your head, that more or less says (means) "I am too tired to have a party on Friday". He also believes that these entities in our heads relate to each other as words and sentences in our normal language, in Fodor's terms, as a 'language of thought'. According to Fodor, the brain (by means of the operations of the language of thought) is able to make inferences and computations on the basis of the content of our beliefs. So, if your head contains the belief that 'all kittens need a lot of attention', and the belief that 'Cleo is a kitten', it will infer and form from those beliefs the third belief that 'Cleo needs a lot of attention'.

Fodor's theory of mind is called the Representational Theory of Mind, or RTM. RTM explains *why* our daily attributions of attitudes are so successful, namely because they simply refer to real entities in the heads of people. Not only that, Fodor believes that it is possible that science will tell us something about attitudes *that we did not know before,* and that we might *change* our attitude talk on the basis of scientific knowledge.

An important reason why Fodor chooses intentional realism is that he thinks only intentional realism is able to prevent what is called 'action-at-a-distance'. Attitude explanations explain behavior on the basis of the *content* or *meaning* of attitudes. If meaning is outside the head (as interpretationists think, as well as many other so-called *externalists*), how could it *explain* or *cause* action? In other words, if it is the meaning of an attitude that causes the action (and it seems that it does, at least in everyday folk psychology), the meaning must be either in the head, or we have action-at-a-distance (Fodor: 1980, cf. Stich: 1991).

Due to meaning holism, interpretationists (and other externalists) indeed have difficulty with action-at-a-distance and with causation more generally. Intentional realists furthermore avoid the problem of indeterminacy of content.

Dennett finds Fodor much too optimistic and his theory of meaning wishful thinking. Even though Dennett thinks that attitudes are, in a certain respect, "real", he does not believe that they have the properties that Fodor thinks they have. Attitudes are no physical entities having causal power, they are not discrete, and they are certainly not in the head. Folk Psychology as a *scientific* theory is extremely limited. It cannot tell us, for instance, whether animals have beliefs, or whether ideas are concepts. Dennett furthermore supports Davidson's thesis that the notion of belief is surrounded by a host of other suspicious problems that make beliefs unappealing as a scientific concept from the start. Is, for instance, the belief that 3 is more than 2 the same as the belief that 2 is less than 3? Useful as folk psychology is, it is not a scientific theory, if only because it is quite unclear how to get authoritative answers to such questions (Dennett: 1987e).

This does not turn the method of ascribing attitudes into a useless theory. On the contrary, it is exactly because the method works that attitudes turn out to have a certain respectable ontology after all. Or at least, that seems to be Dennett's claim.

The debate over intentionalism realism and the question how and whether one should account for the causal power of intentional states is usually seen as an ontological debate. Whoever wants to take seriously intentional explanations, will have to face deep ontological problems like the problem of mental causation, and the mind-body problem (the core of these problems is well laid out in Kim: 1989; Kim: 1993; Kim: 2002). I by and large disregard these ontological problems in this thesis and instead interpret Dennett's theory of the stances as an *epistemological* theory. After all, Dennett himself proposes to see his theory as an *epistemological* theory. The theory of the stances is not about attitudes themselves; it is about us ascribing attitudes and about us correctly ascribing attitudes.

I will give a detailed account of Dennett's theory of the stances as an epistemic theory in the next chapter. But let me here shortly sketch what I mean by an epistemological approach. The epistemological or explanatory approach starts with the claim that priority must be given to *explanations*, that is the epistemic situation (for Dennett: the stances). Ontology, then, follows epistemology. I take Quine and Philip Kitcher (Kitcher: 1985) to be the clearest defenders of this idea, and Dennett to be an implicit follower of the idea.

Quine developed this idea in the notion of 'ontological commitment': we are ontologically committed to those entities that figure in our best explanations of the world as we experience (talk about) it. Quine, of course, did not think that intentional explanations are good explanations – and therefore never really had to deal with problem of mental causation anyway (we can regard Quine to be an eliminativist). But the Quinean idea of ontological commitment has been adopted by non-reductionists (defenders of the use of higher-order explanations in science), most notably by Kitcher.

I read Kitcher's revision of the notion of causes as an attempt to extend the epistemological approach to the domain of higher-order explanations, notably intentional and functional explanations. Kitcher claims, rather like Quine, that explanations should come first. If we can find epistemological reasons for accepting a certain (higher-order) explanatory paradigm (in Kitcher's case: functional explanations), we should not concern ourselves too much with the ontological problems that may come with it.

Sciences use many kinds of explanations (higher-order explanations, functional explanations, intentional explanations) that improve our understanding of the world (see, for instance, Jackson and Pettit: 2004, who argue for a distinction between program explanations and process explanations). These alternative explanations cannot and should not be dismissed so quickly. If the problem of mental causation forces us to abandon these alternative explanations, while science relies so heavily on them, then we could just as well conclude that there must be something wrong with the ontology. Kitcher, amongst others, has therefore suggested that primacy should be given to good explanations, not to causation. The ontic conception of explanation needs to be abandoned in favor of an epistemic theory of explanation (see Meyering: 2000, 200, for a nice summary of Kitcher's position; see also Baker: 1995 for a slightly different, but comparable approach).

A consequence of the epistemic approach is that it cannot make strong ontological claims anymore, in particular with regards to causation. 'Causal' explanations derive their value from the fact that they can – ideally - give precise and full predictions of future events – not from the fact that they refer to "real causes", or "realest causes" in the world.

I believe that Dennett holds a Quinean idea of ontological commitment – quite similar to Kitcher's revised notion of causation. In *Real Patterns* (1999) Dennett indeed suggests that he might support such a view and opt for a more idealistic notion of causation: If one has found a predictive pattern, one has *ipso facto* discovered a causal power (Elton: 2003, 96). Attitudes then are, due to their indispensable role in Intentional Systems Theory, part of our ontology. Not as physical entities ("illata"), but as theoretical constructs ("abstracta"). Ontologically, this is a rather suspect position, because our ontology would contain both physical and intentional elements and who knows what else. The "world" would easily become overpopulated and messy – we could speak of a "jungle ontology" or what Ross has called, again after Quine, "rainforest realism" (Ross: 2000). Yet, an interpretation like this has been proposed by several Dennett interpreters and seems to be very plausible (see, for further details on such a view, Viger: 2000 and Ross: 2000). Dennett himself has agreed that if he has to opt for an ontology, he might indeed favor some kind of rainforest realism (2000).

As said, I regard this thesis as a thesis in epistemology, not in ontology, and I take Dennett to defend an epistemological theory, not an ontological one. Were Dennett to defend an ontology, it would probably be around the contours of the approach described above. What matters for me at this point is that Dennett believes that intentional explanations could be genuinely good explanations, *and* that he needs to show that this is so in order to get his whole project started. What also matters is that Dennett rejects the common ontology of attitude ascriptions, but not the methodology. Intentional explanations need not refer to entities that look similar to those that make up our everyday ontology – but they do need to be predictive and explanatory in order to be justified.

Dennett's epistemic approach to attitudes is also characteristic for his pragmatic naturalism. Deep ontological problems about the mind are put aside, and those features that are usable and practical, such as explanatory power, put to the front. Perhaps this is cheating, but it is very consistent with the main pillars of the theory.

3.7. Two standards

This chapter served to lay out the essential elements of Dennett's theory, and to get some grip on what Dennett himself considers or should consider to be basic standards for evaluating a theory of ascription. I am now in the position to formulate such standards.

For as far as Dennett cares to deal with ontological issues at all, he claims that the ontology that is presupposed by our ordinary attitude ascriptions is misguided. Analyzing the concept of an attitude on the basis of our ordinary attitude ascriptions will give us the way we use attitudes terminology in daily life, which may wrongly lead us to the idea that we have, in doing that, said something about *attitudes themselves*. This is wrong, Dennett thinks; our daily theory of attitudes is largely wrong in an ontological respect: there are no such things as beliefs in the head, discrete entities with a specific content that interact and as such cause us to act. For these reasons, Dennett does not think that science will eventually vindicate our everyday theory of attitudes – in *this* sense.

That is not to say that Dennett does not care about what attitudes are. As a matter of fact, Dennett's starts his philosophy of attitudes with a descriptive account: a description of how we take the intentional stance in daily life, how intentional beings evolved, and the characteristics of intentional states and the stance. Contrary to Fodor, he believes that its justification lies not in ontological validation of the referents of attitude ascriptions, but in the practical virtues of the intentional stance itself.

By making this switch, Dennett can also provide an answer to the normative question: to what extend should we take our intentional ascriptions seriously? It is not coincidental that, for Dennett, the actual way of ascribing attitudes is more or less similar to the *right* way of ascribing attitudes in the methodological sense. As a matter of fact, it is mainly due to the explanatory and predictive strength of the intentional strategy that the intentional stance is worth evaluating as a method for science. A good method, for a pragmatist and a naturalist like Dennett, at least has great predictive powers and great explanatory strength. These two features come cheap with the intentional stance. That we take the intentional stance so often, an ability that has grown on us during our recent evolutionary history, and that we are therefore able to quickly predict and explain intentional events with it is a prima facie reason to think that it may also be a good method.

Whether it can indeed serve as a good stance in *science* is then still to be shown of course (see chapter 7 for an argument that it can). I should note at this point that it need not *necessarily* be a *scientific* method. It may also be a conceptual method, for instance. Recall the discussion about personhood: what mattered to Dennett in that case was not what a person is from a scientific point of view, but to devise a concept of personhood that can help us answer certain questions.

Ideally, then, the descriptive take and the normative take of the intentional stance come together for a pragmatic naturalist like Dennett. Both are important in evaluating a theory of ascription and ideally they are in harmony.

Dennett indeed seems to want to connect the descriptive (naturalism) and the normative (pragmatism) in this way. Dennett describes intentional regularities as patterns, real patterns, in an attempt to show his critics that his position is not straightforwardly unrealistic (Dennett: 1999). The intentional stance discerns these intentional patterns:

> "I claim that the intentional stance provides a vantage point for discerning (...) *useful* patterns. These patterns are objective – *they are there* to be detected – but from our point of view they are not out there entirely independent of us, since they are patterns composed partly of our own "subjective" reactions to what is out there; they are the patterns made to order for our narcissistic concerns (...). *It is easy for us, constituted as we are, to perceive the patterns* that are visible from the intentional stance – and only from that stance." (Dennett: 1987c, 39)

Patterns are, as Dennett calls them, recognizabilia. Even though they are only perceivable while adopting a certain stance (the intentional stance in the mental case), and thus for those who can take this stance, Dennett thinks these patterns exist even when we would not actually perceive them. Hence: recognizabilia: things that are, in principle, recognizable. Thus, the 'right' ("real", "useful") way of seeing intentional patterns as a matter of fact is for Dennett continuous with the actual way we see intentional patterns.

The way Dennett joins the descriptive and the normative elements of his theory of attitudes shows very nicely how his naturalism and his pragmatism come together. Naturalism requires that a theory is firmly embedded in the empirical facts: a theory of ascription should be descriptively correct. Pragmatism requires a theory of ascription to have some practical value; in Dennett's case, methodological value, or, as I prefer to more specifically call it, methodological utility. These two requirements should be complementary. Normative claims are suspicious in every naturalist theory, so the normative theory of ascription better be continuous with a descriptive theory of ascription, if not ontologically, then qua method. As a full blown naturalist, Dennett surely would not want to go beyond the empirical facts more than he has to. And why should he: it is the intentional stance as we *in fact* take it that has the remarkable explanatory power that we know it to have, and it is thus the intentional stance as we in fact take it that bears the methodological utility that Dennett seeks.

I thus propose the following two standards that enforce themselves upon the naturalistic-pragmatic approach that Dennett wants to develop:

(1) *empirical correctness*: a theory of ascription should be descriptively correct (a theory of ascription of *x* should say something about how we actually ascribe *x*'s)

(2) *methodological utility*: a theory of ascription should be methodologically practical (a theory of ascription of *x* should help us attain a goal, usually a scientific goal)

At this point I keep these standards rather vague. Their exact contents will become clearer in the chapters to come, when I deal with them in more concrete detail.

4 Interpretationism as Non-Reductionism

This chapter takes an excursus into the more technical details of Dennett's theory of the stances. Those who only want to follow the main argument of the dissertation, can skip this chapter.

Dennett's pragmatic outlook on intentionality may make him appear to be a straight reductionist. This is not at all the case, I argue in this chapter. I interpret Dennett as a (epistemic) non-reductionist, and place him in the camp of functionalistic non-reductionists that see intentional explanations as a kind of functional explanations.

This chapter is multi-functional. It serves as a more specialized chapter for clarification of Dennett's theory (or at least my interpretation of it) and as a place of reference for certain more technical terms that may cause confusion for those working on the technical details of Dennett's theory. It also relieves chapter 7 from a reductionist objection that a methodological account of the stances cannot even start to be successful.

Dennett's pragmatic take on the mind and on intentional explanations easily gives the impression that his theory of the stances should be read as basically a reductionist thesis about intentional explanations. Indeed, many read Dennett as defending intentional explanations *merely* as a useful heuristic. Partly, this impression is fueled by a persistent confusion of ontological and epistemological reductionism[33]. Ontologically, Dennett's theory may well come down to be a reductionist thesis (like almost every monist theory of the mind in the analytic world today), but the relevant frame of reference for my thesis is epistemology. From this perspective Dennett's account can be construed in a much richer way. The intentional stance (and the design stance) turn out to be explanations in their own right, with their own powers and weaknesses, that may do good service even as a scientific method (as I concrete develop further in chapter 7).

[33] As mentioned, ontology and epistemology often are not very well distinguished in the literature, and it is not easy to tear them apart. It is not always possible to keep an epistemological discussion purely epistemological, even if you choose your words carefully. In addition, there are certainly well-defended ontological (e.g. Quinean) positions in which ontology reduces smoothly to (scientific) epistemology. For such positions, the distinction evaporates altogether (see 3.6).

Those who want to defend the use of intentional explanations in science, like Dennett wants to, will usually try to show that these explanations provide something that cannot be had by explanations of a lower order. At least, and more in particular, they should be (1) *valuable* explanations that are (2) *irreducible* to lower level explanations. Typically, it is shown that they are irreducible *because* they are valuable as higher order explanations. Dennett, as I interpret him, justifies his account of the intentional stance with this non-reductionist strategy.

I first discuss in a general way Dennett's ideas about how intentional explanations relate to lower level (usually physical) explanations (4.1). I then give a short characterization of non-reductive explanations. By construing different strengths of (non)reductionism in contemporary philosophy of science (4.2), I will be able to place Dennett on a reductionism-scale, concluding that his theory of the intentional stance is in fact a quite strong form of non-reductionism (4.3).

4.1. Intentional systems Theory and Subpersonal Psychology

Dennett makes a rather sharp distinction between intentional explanations (explanations at what is called a 'higher' level) and reductive explanations (e.g. physical explanations on a lower or the lowest level). In *Three kinds of Intentional Psychology* (1987e), Dennett explains this best.

As discussed in chapter 3, Dennett does not think that Folk Psychology can be taken as a literal scientific theory, but he does believe that it can play a methodological role in science. In order to show this, he makes a strict division between two separate scientific frameworks: Intentional Systems Theory (IST) and Subpersonal Cognitive Psychology (SCP).

Dennett relates the distinction between IST and SCP to the Rylean distinction between conceptual answers and causal answers (see 3.3). When we ask what some *x* has in common with all other *x*'s, we can give both a conceptual and a causal answer. For instance, when we ask what all magnets have in common, we can give the generalizing answer that they all attract iron. We are then giving a general definition, a theoretical answer – often dispositional. The causal answer, by contrast, gives a reductive answer, which explains *why* all magnets attract iron in terms of their physical properties. According to Dennett, these two types of answers do not rule each other out. Conceptual answers about the mental, for instance, cannot be given by microreductive psychology, and conceptual answers cannot explain anything about the causal workings of the mental. Stronger even, to try to explain one in terms of the other would be making a *Rylean category mistake* (Ryle: 1949, cf. Dennett: 1987e, 44, see also Dennett: 1969). Both types of answers are good answers. (45)

I'm not particularly happy with Dennett's use of the term 'conceptual'. It may wrongly be associated with conceptual analysis. Furthermore, in the rest of his work, the idea that intentional explanations are *conceptual* gets no further development or explanation. What matters here, however, is that Dennett wants to distinguish between two *types* of explanations that give different *types* of answers.

What, then, do the two types of frameworks explain? *Intentional Systems Theory* is a theory about intentional states and rational agents, where attitudes are seen as idealized phenomena that are understood holistically. They derive their legitimacy instrumentally, by being good, useful theoretical concepts within a scientific framework. Dennett likes to compare attitudes with *abstracta* from the physical sciences. An example of an abstracta is a centre of gravity. Abstracta are good and useful means that help us explain and predict the world, even though we cannot trip over them (they are abstract) and despite their instrumental nature. Likewise, notions such as 'belief' and 'desire' are useful abstract notions – immaterial and abstract, which can, in principle, be perfectly legitimate in a scientific enterprise (see also {Dennett, 1999 2 /id).

The intentional stance can and is used in decision theory, rational choice theory, game theory, or any other theory that deals with the explanation and prediction of rational agency. Such theories all pose some idealized agent; fully rational or endowed with some form of bounded rationality. This is a methodological assumption (methodological rationalism) – no rational choice theorist is committed to the claim that real people *are* fully rational, only that people's behavior can be explained and predicted by assuming that they are.

Intentional Systems Theory can, according to Dennett, also play an important role in evolutionary biology if we want to capture the intentional commonalities of a certain species, commonalities that have contributed to their survival value (examples will be discussed in 7.2).

Subpersonal Cognitive Psychology, on the other hand, studies the concrete, *micro-physical realizations* of intentional systems. SCP gives the reductive, 'causal answer', by means of neurochemical or neurophysiological theories, etc. Whenever we explain the behavior of an intentional system, we have a choice: *interpret* its actions from the intentional stance (IST), or *analyze* the system and study its concrete realization. In the first case we give generalizing, *intentional explanations*. We explain intentional behavior *in terms of beliefs*, desires and so on. In the second case we give causal, reductive answers of a system *to which we can ascribe intentional states*. Importantly: we are then not explaining intentional behavior anymore:

> [T]he cognitive psychologist cannot ignore the fact that it is the realization of an intentional system he is studying on pain of abandoning semantic interpretation and hence psychology" (64)

IST and SCP thus approach intentionally describable phenomena in different ways. One might say that they pick out different phenomena (beliefs versus brain states), or are based on different scientific strategies (e.g. abstracta versus illata); their subject matter is different. IST is about whole systems, usually persons (if we talk about human beings) and is able to capture semantics, SCP is about the 'subpersonal'- brain states, and can only say something about syntax (Dennett: 1987e, 59).

Dennett thinks that it will turn out to be impossible to neatly map intentional states (the abstracta cited in our intentional explanations) onto neurophysiological items. The 'whole' (personal) cannot be neatly analyzed into its parts (subpersonal). Consider the counterfactual "Had Tom not believed that *p* and wanted that *q*, he would not have done *A*":

> "Tom was in some one of an *indefinitely* large number of structurally different states of type *B* that have in common just that each of them *licenses attribution* of belief that *p* and desire that *q in virtue of its normal relations with many other states* of Tom, and this state, whichever one it was, was causally sufficient, given the "background conditions" of course, to initiate the intention to perform *A*, and thereupon *A* was performed, and had he not been in one of those indefinitely many type *B* states, he would not have done *A*" (Dennett: 1987e, 57, my italics)

Thus, intentional states are only reducible to an *infinite* disjunction of physical facts, whereby this disjunction is necessarily held together in terms of belief *attribution* (and, thus, not in physical terms). Moreover, as we saw, such belief attributions can only be made by taking into account relational or holistic aspects. Intentional explanations, then, study behavioral dispositions that underlie intentional regularities. They do *not (have to)* explain or refer to the underlying mechanisms realizing those dispositions (see Viger: 2000, 136). In our explanatory practices, however, the two approaches will often be related. The study of an intelligent system will usually use both explanatory strategies simultaneously.

Obviously, for a physicalist, everything in the world, *in the end*, reduces to physical reality. But this is an ontological claim. What matters is whether we *should* aim to reduce everything to the physical level.

As I read Dennett, his position amounts to the following. Yes, we can reduce all phenomena in the world to some lower level, and yes, sometimes we learn a lot from it, *but*:

(1) Science does not tell us that we always *have to* do that. In fact, scientific practice uses abstractions (models, maps: abstracta) all the time and clearly in a respectable way (think about centers of gravity). Science is in principle free to use whatever explanatory level it sees fit for its epistemic goals.

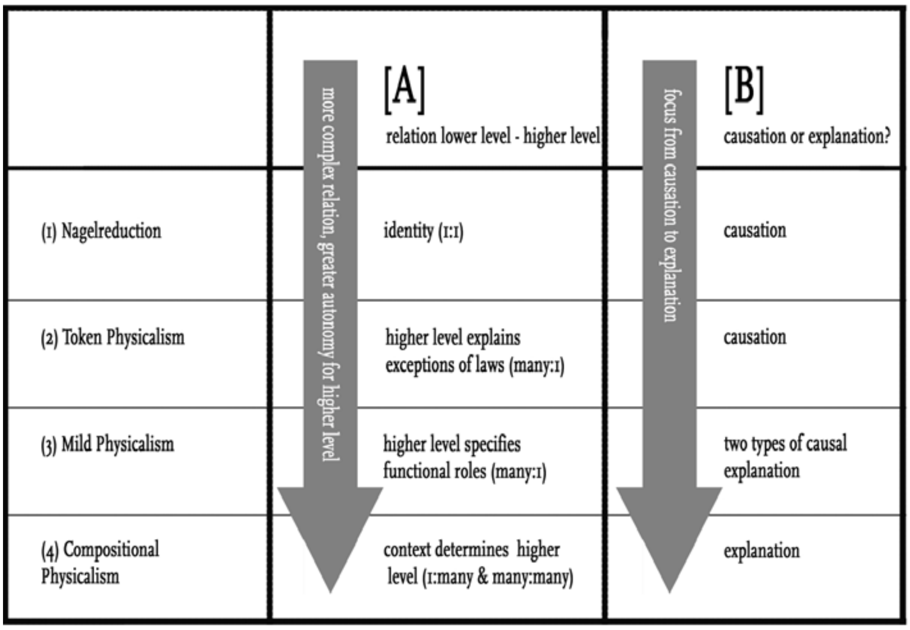
(2) In the case of psychology, we can *only* get a reduction going by working from a different framework, SCP, and *only* on pain of ignoring intentional patterns. Whenever we are reducing intentional behavior to, for instance, physics, we are not talking about attitudes anymore, and neither are we explaining them. If we want to explain attitudes, or use them as abstracta in our scientific theories, we will have to use IST. There is, as such, a distinction to be made between intentional explanations (IST) and the explanation of intentions (SCP). It is not plausible that we will (any time soon) have a workable psychology devoid of intentional terminology.

4.2. The reductionism scale

Intentional explanations are suspicious in science. To account for this, many philosophers of science and psychologists defend some kind of use of intentional explanations and other "higher-order" explanations in the (psychological) sciences. Usually, this is done by means of a non-reductionist argument.

In this section I will distinguish several ways to be a reductionist and order them by means of answers to two questions. The first question is about the relation between lower and higher levels, or, about *intertheoretical relations*

(McCauley: 1996, 28). As I will explain below, a reductionist is at least committed to the claim that there is a traceable and intelligible link between predicates of the special sciences and those of the physical (or lowest level) sciences (*A*). The non-reductionist typically denies this. The second question is whether or not only physical science, or lower-level sciences, can offer true, (i.e. *causal*) explanations (*B*). Non-reductionists will also tend to answer the second question negatively. This leads to the 'reductionism scale' presented in Figure 1 (based on the very helpful overview made in Meyering: 2000).



	[A] relation lower level - higher level	[B] causation or explanation?
(1) Nagelreduction	identity (1:1)	causation
(2) Token Physicalism	higher level explains exceptions of laws (many:1)	causation
(3) Mild Physicalism	higher level specifies functional roles (many:1)	two types of causal explanation
(4) Compositional Physicalism	context determines higher level (1:many & many:many)	explanation

more complex relation, greater autonomy for higher level
focus from causation to explanation

Figure 1: The reductionism scale

I will explain this table in detail shortly, but let me first explain its structure. In the left column, we see four main positions in the reductionism debate, from reductionist positions (above) to non-reductionist positions (below). The lower in the table, the stronger the form of non-reductionism. The positions on the scale correspond with the type of answers it gives to the two main questions, (A) whether it is possible to smoothly map levels onto each other (a 1:1 relation is smooth, whereas a many-many relation makes mapping very hard or even impossible); (B) the role of causes in the explanations of higher order sciences (the non-reductionist typically settles for a more liberal notion of causation).

1. Nagel-reduction.

Let's start with reductionism. Strong versions of reductionism, which I call Nagel-reduction (after Ernest Nagel[34]), say that laws of the special sciences can and should be reduced to physical laws by means of the identification of special kinds with physical kinds. The predicates that play the role of antecedent and consequent in the special law are to be *identified* with their physical correspondents. So-called bridge laws then enable the reductionist to translate the special law into a law of physics.

Nagel reductionism is nicely illustrated by this example of Doug Snodgrass[35]:

SPECIAL LAW:

(1) S1x → S2y (S is depressed → S has thoughts of worthlessness.)

BRIDGE LAWS:

(2) S1x ↔ P1x (S is depressed ↔ S has serotonin imbalance.)

(2b) S2y ↔ P2y (S has thoughts of worthlessness ↔ S has patterns of neural firings in the brain such that ______.)

PHYSICAL LAW AFTER REDUCTION:

(3) P1x → P2y (S has serotonin imbalance → S has patterns of neural firings in the brain such that ______.)

As the relation between special kinds and physical kinds must be one of *identity*, Nagel-reductionism is strongly tied to the identity theory.

One important reason to be a Nagel-reductionist is that the position is immune to the problem of downward causation. The problem of downward causation was well described by Kim for the realm of the philosophy of mind (Kim: 1989[36]):

[34] Few philosophers support the strong version of reductionism as advanced by Ernest Nagel. But the identity theory is still alive and kicking. I count the identity theory as a somewhat mild version of Nagel-reductionism.

[35] http://www.dasnodgrass.com/college_life/philosophy/On_Fodor_Special_Sciences.pdf

[36] Kim changed his ideas about mental causation in his more recent *Mind in a Physical World* (Kim: 1998). Nevertheless, his explanation is the best for my purposes because it is simple and adequate, so I will use it in his name in this paper.

(assumption 1) If you do not want to be an eliminativist – i.e. recognize the existence of mental events – you better make sure mental properties have causal powers

(assumption 2) Physicalists have to assume that the physical world is causally closed (causal happenings in the physical domain always have a physical cause).

Now, what if a mental event causes a physical event? What, then, is the relation between the mental (say, pain) and the physical (retreating the hand)? Or, what caused my hand to retreat? Partial causation (both the mental and the physical are causes) and overdetermination (both the mental and physical are sufficient causes) are both implausible (44). There is, then, according to Kim, only one solution: the mental and the physical are identical to each other. The argument against downward causation is thus a strong argument for reductionism. According to Kim, there really are only two options if you don't want to be a reductionist. Either be an eliminativist (deny assumption 1), or be a non-physicalist (deny assumption 2). Kim thinks this exhausts the options for the non-reductionist and that the main position in the philosophy of mind today, non-reductionist physicalism, is inherently unstable.

For the Nagel-reductionist, the relation between levels is that of identity. As such, Dennett's account can never be Nagel-reductionist. For Dennett denies that we can make such direct translations. For the Nagel-reductionist, causality takes place at the physical level[37]. Non-reductionists have to reply to the strong intuition that the real causal work indeed takes place at the lower level – or they have to deal with the problem of downward causation in some other way. The non-reductionist will at least have to show how higher-order explanations can be good explanations, either by showing that they refer to real causal relations at this level, or by altering the notion of 'good explanations'. The non-reductionist also has to reject the claim that special kinds can be identified with physical kinds. Let me discuss some options.

[37] Although causality can be taken in an epistemological way, Nagel clearly gave it a stronger, ontological meaning. I take it that Nagel-reductionism can – in principle – be interpreted as an epistemological position; even if Nagel himself did not take it as such.

Heurism

Before I start out to discuss a number of important non-reductionist theses, I should mention a position one does not find on my scale: heurism.

'Heurism' is my term and refers to all theories that point at the instrumental, practical, or, indeed, heuristic value of special kinds and higher order regularities. Usually, this kind of thesis is interpreted as a reductionist thesis: higher order predicates are useful, but they are fictions, reflecting temporary gaps in our knowledge. *I* take it that heurism in principle fits with whatever approach you take with respect to the reductionism debate. Of course, those who claim that higher order entities have *only* heuristic or instrumental value, and *only* with respect to the reductionist project, are committed to some form of reductionism. At the same time, and consequently, I do not think that heurism alone can establish a proper reductionist thesis.

This will prove to be important, because some think that Dennett's interpretationism amounts to nothing more than simple heurism and conclude that Dennett is a reductionist. As we will see, Dennett thinks that higher order regularities *are* real relative to their usefulness to us, but he does not think that they can be used for some reductionist project. As such, I claim, Dennett's heurism does not amount to reductionism.

2. *Token Physicalism* is famously defended by Jerry Fodor, and is meant as an alternative to the strong demands of Nagel-reductionism, while keeping intact the ideal of the unity of science. In his important paper 'Special Sciences' (Fodor: 1974), Fodor argues that Nagel-reduction is impossible, and uses the functionalist multiple realizability claim as his main argument (see also Putnam: 1975). Special kinds are, according to Fodor, *wildly disjunctive* at the level of their physical implementation. A financial transaction, for instance, can be instantiated by a coin OR a dollar bill OR an online transfer... etcetera. Moreover, the disjunction should also include all possible future instantiations, in order for the reduced physical law to support the relevant counterfactuals. Similarly, minds or beliefs are multiply realized as well (dolphins and Martians may also have minds).

According to Fodor, the multiple realizability argument is a major problem for Nagel reductionism. We would have to identify special kinds or laws with these wild, possibly infinite, disjunctions. Fodor thinks such disjunctions could hardly figure as a predicate in a (physical) *law*. Moreover, even if it were possible to translate the special kinds into such a disjunction, it is quite unlikely that the elements of this disjunction would have anything more in common than their special description. And, *whether* the physical description of special events or things have anything in common is often irrelevant. Fodor says: "[W]hat is interesting about monetary exchanges is surely not their commonalities under *physical* description" (134).

Let me notice that so far, Dennett would completely agree with Fodor. But Dennett would presumably be less optimistic about Fodor's solution. According to Fodor:

> [t]he point of reduction is *not* primarily to find some natural kind predicate of physics coextensive with each kind predicate of a special science. It is, rather, to explicate the physical mechanisms whereby events conform to the laws of the special sciences" (138)

In other words, Fodor thinks we can and should map the ways in which special laws are instantiated - especially because special laws have exceptions (and as such are not really laws) that need to be explained from the lower level. So, Token Physicalism denies that type-type identification of special kinds with physical kinds is possible, but does believe in the microanalysis of the tokens. Hence: *token-physicalism*. As we saw, Dennett thinks such mapping will most probably be impossible, and is therefore not a token-physicalist.

Although Fodor's position is an alternative to Nagel-reductionism (reducing types to types), and although his arguments are usually taken as a hall-mark argument against reductionism, it is still a rather mild form of reductionism that says that we can and should understand higher order regularities in terms of their physical realizers at the *token level* (cf. Meyering: 2000).

3. Mild Physicalism is, according to Meyering (and I agree), the first real candidate truly worth the name non-reductionism. The argument from Multiple Realizability is used in this context to show that higher order states can reflect functional *roles*. Mild physicalists claim that the regularities that hold at a higher level can be explanatory by themselves, and, importantly, that explanations at this level do not necessarily conflict with explanations at a lower level. Frank Jackson and Philip Pettit (Jackson and Pettit: 2004), for instance, distinguish *process explanations* (citing the actual causal agents - realizers) from *program explanations* (citing function, role, or general nature). By distinguishing types of explanation, purportedly non-conflicting, this position tries to shake off the Downward Causation problem in functionalism. Mild Physicalism claims that sometimes only the higher order explanation sustains the relevant counterfactuals[38]. Meyering uses as an example the event that a stampeding herd crushes a little boy's toy. When true, it seems hardly causally relevant that a particular cow actually crushed the toy - some other cow would have done it if this particular one hadn't (190). So Mild Physicalists take both types of explanations to be causal explanations.

Mild physicalism is a (and the first) truly epistemological position; ontological considerations are of later concern. As I will discuss later, Dennett's justification of intentional explanations comes quite close to this position.

4. *Compositional Physicalism* denies, contrary to previous positions, that special events supervene unilaterally on physical events on the basis of the idea of Multiple Supervenience[39]. Multiple Supervenience says that some physical item or event can instantiate multiple higher order events. Baker (Baker: 2000), for instance, argues that a certain piece of marble can be just that – a piece of marble – but that it can also be Michelangelo's David, depending on the relations of the marble with its environment. Similarly, a big rock can be either a planet or a mountain (or...), depending on its context. Multiple Supervenience, or constitution, is exactly the opposite of Multiple Realization, which says that some special kind or event could be realized by multiple physical items or event, e.g. a hammer can be made of wood or made of plastic.

The argument from Multiple Supervenience is a strong argument against so-called structure-or species dependent reductionism. Many reductionists today, at least in the philosophy of mind, admit that psychological kinds in general can be multiply realized, but believe that there is a one-to-one relation between human beliefs and human brains (cf. Kim: 1989, 38-39). Those who believe in multiple supervenience could attack precisely this claim. What higher order properties are instantiated in a certain brain depends, according to those non-reductionists, on

[38] Compositional physicalists presumably also support this claim

[39] I take it that multiple supervenience accounts amount to the same position on the reductionism scale as constitution accounts, as used by Baker (2000). This is not to say that the positions are equivalent, of course.

the context[40]. Furthermore, it is difficult to see how we would, for instance, distinguish the class of believers as a whole, without the notion of a believer itself (an argument from Dennett that will return later).

A variant of compositional physicalism is the systems approach. Whereas the constitution position emphasized the relation of some item with its environment, system approaches focus on the whole system itself, or its *systems properties* (see, for instance, Bechtel and Richardson: 1992).

Most Constitutional Physicalists take higher order explanations to be truly causal explanations. Baker, for instance, thinks intentional explanations are causal explanations, but denies that the attitudes are brain states (Baker: 1995, 28). Explanation then becomes prior to causation; the 'because' of causation in fact is quite close to the 'because' of explanation.

Kitcher, most prominently, argues that the problem of mental causation lays too much emphasis on causality in the physical realm, and too little on explanation. It assumes that the only good explanation is an explanation that traces real causal patterns, where real causal patterns are physical patterns: strong causal realism. This seems contrary to normal scientific practice: science uses many kinds of explanations (higher-order explanations, functional explanations, intentional explanations, program explanations), and these explanations improve our understanding of the world. If Descartes' problem forces us to abandon these alternative explanations while science relies so heavily on them, then there must be something wrong with the ontology. Or so the argument goes. Kitcher, amongst others, has therefore suggested that primacy should be given to good explanations, not to causation. The ontic conception of explanation needs to be abandoned in favor of an epistemic theory of explanation. For Kitcher, unification and systematization of beliefs should be our main concern (see Kitcher: 1984, see also Meyering: 2000, 200, Baker: 1995, Pettit: 1993).

Compositional Physicalism shakes off the problem of downward causation (or rather, happily embraces it), but has problems of its own. Not only does it easily

[40] Species- or structure-reductionism is, I think, problematic for other reasons as well. It is hard to see how such a form of reductionism would work, for instance, for financial transactions. Would we then have a science of cash-transactions, and a science of cheque-transactions and…?

lead to unattractive forms of subjectivism or relativism, but it is also unclear how to distinguish real causal processes (whatever those may be) from pseudo-processes. This approach then needs to show how good explanations can be distinguished from bad ones[41].

4.3. The intentional stance as non-reductive explanation

I believe that, seen from the epistemological reductionism scale I just presented, Dennett's interpretationism is a strong form of non-reductionism. One might argue, against me, that Dennett is an in-principle reductionist because Dennett believes that under conditions of ideal Laplacian science we *could* predict every event in the world from the physical stance. But we should keep in mind that there is a difference between reductionism and *physicalism*. Physicalism is an ontological claim and says that ultimately everything in the world is physical[42]. A non-reductive physicalist, then, claims that there are nevertheless good reasons to accept non-reductive explanations in science. Most non-reductionists today are *epistemic* non-reductionists. (for a modern version of ontological non-reductionism, see for instance Baker: 2000). Seen from the epistemic perspec-

[41] Perhaps one would expect a sixth position on the reductionism scale: dualism and phenomenological approaches that claim to be non-dualists. But dualists are typically *ontological* dualists, or ontological non-reductionists, and are therefore not part of my epistemological scale. Ontological dualism would amount to saying that there are different kinds of stuff in the world that lead their own lives (have their own causal effects). The kinds of non-reductionism I have dealt with are all *explanatory* forms of non-reductionism – forms of explanation that could be adopted by science. And this is exactly what the dualist argues against. Thomas Nagel, for instance, famously claimed that: "an organism has conscious mental states if and only if there is something that it is like to be that organism - something it is like for the organism." (Nagel: 1974, 436). And for Searle, first person statements are "made true by the existence of an actual fact that is not dependent on any stance, attitudes, or opinions of observers" (Searle: 1996, 9).

[42] Again, the distinction between ontology and epistemology is often very hard to make. Take genes for example: do they have ontological integrity or not? Thinking in terms of genes helps us explain a lot, but technically they are merely specific bundles of molecules (Waters: 1990, Gasper: 1992, Rosenberg: 1997). We can take the non-reductionist to say that ones ontological views on what are the final building blocks of the world are relatively unimportant. Even if there are, in the en, only molecules, atoms, quarks or strings, there may still be reasons to accept higher order compositions as objects to study, and higher order explanations to explain the behavior of these objects.

tive, Dennett's interpretationism falls on the far end of the non-reductionist positions. Why?

We saw that the positions on the reductionism-scale are determined by two factors:

(1) *the relation between levels.* Reductionists claim that one level can be smoothly mapped onto a lower level (or else: the level should be eliminated). Non-reductionists deny this. The less intertheoretic mapping is possible, the more a theorist will tend (or be forced) to say that higher level entities are not reducible to lower levels (cf. McCauley: 1996).

(2) *explanation and causation.* Reductionists might claim that higher order events explain, in a provisional sense, something about the world, but insist that real explanation takes place at the physical level. Non-reductionists, however, will have to show why higher order explanations are explanations proper, either by revising their notion of causal explanation, or by showing that higher-order explanations can peacefully co-exist with physical or causal explanations.

With respect to the *relation between levels,* Dennett's position is clearly that there is no smooth translation of higher order (*functional/ intentional*) predicates to physical predicates. At the very best, a translation would yield an infinite disjunction of physical particulars – Fodor's position. But Dennett has a much stronger position than Fodor. Fodor claims that we can (and should) in principle find a physical description of every *particular* higher order event, and isolate its causal powers. Dennett does not believe that that is possible, or rather: at least not *always* and certainly not in the case of mental events in current science (it worked for chemistry, but chemical properties are still real patterns). Intentional patterns, at best, supervene globally on the physical. Furthermore, an explanation of such events in terms of its underlying mechanisms would, according to Dennett, be irrelevant to the truth, explanatory value, or even reality of higher order events.

Dennett seems to be, at least, in line with Mild Physicalism. Dennett's distinction between conceptual and causal answers (and, relatedly, explanations that refer to illata and those that refer to abstracta), seems to me to be equivalent to the distinction Jackson and Pettit make between program explanations and process explanations. Dennett would agree with Jackson and Pettit that such explanations are of different types, that they do not have to conflict, and that program explanations cite exactly shared (functional) properties.

I think that Dennett would also support the idea of systems explanations as defended by Compositional Physicalism. The notion of a belief is for Dennett clearly a notion that is attached to a *system*, and to systems only (namely: intentional systems, and intentional systems only). Whether Dennett would also support the idea of multiple supervenience in this context, however, is more difficult to answer. The multiple supervenience thesis, at least in the form Meyering defends it, suggests that we can isolate lower-order entities and determine what they "are", or their causal effects, by studying their relational properties (Meyering speaks of the causal relevance of relational properties). If I have interpreted Dennett correctly, he does not think such a link between levels is possible in the case of intentional events. The physical level and the intentional level require different stances, and the reality of the latter is not determined by the reality of the former. His holism of intentional states, though, does take seriously the idea that beliefs can only be understood as relational properties, i.e., relations with the environment, and with other intentional states (a view he shares with Baker: 1987). So, although it is unclear whether Dennett would adopt the methodology that is attached to the multiple supervenience argument, the basic model seems pretty much the same.

What are Dennett's views on *causality and explanation*? Elton (2003) thinks that Dennett's view pulls in two directions. In *Three Kinds* (2002b), Dennett seems to *deny* that intentional states have causal powers. In *Real Patterns* (1999), however, Dennett suggests that to deny that intentional states have causal powers is a mistake, based on a simplistic notion of causation: if one has found a predictive pattern, one has ipso facto discovered a causal power (Elton: 2003, 96). In the first case, the physical has causal powers, whereas the intentional has not. In the second case, both have causal powers, where 'causality' is understood in a broader sense - in line with the strong versions of non-reductionism I discussed above.

Elton thinks Dennett should opt for the first option: deny that intentional states have causal powers. This seems reasonable, given Dennett's distinction between causal and conceptual answers. Christopher Viger, another Dennett interpreter, defends the second reading (Viger: 2000). Viger thinks Dennett wants to get rid of the whole dichotomy of causal/real/physical explanations at the one hand, and instrumental/unreal/intentional explanations at the other. In any case, Dennett does not hold the traditional notion of causation as the reductionist wants to have it.

4.4. Conclusion

Dennett's theory is often put aside as simple reductionism, but as a matter of fact it should be seen as a form of epistemic non-reductionism, bearing many characteristics (liberal notion of causality, incommensurability of levels etc.) of epistemic non-reductionism that have been and are defended by many philosophers of science.

I emphasize this because it has been held against Dennett a lot that his theory is ontologically dubious. By framing him as an epistemic non-reductionist, this criticism can be met by pointing at the fact that the same ontological problems would infect all other non-reductivist theories and are, hence, general problems and not specific to Dennett's theory. In the worst case scenario that such criticism cannot be met, epistemic non-reductionists can always bite the bullet and admit that their theory is in an *ontological* sense reductivist. Most non-reductivists won't find this an unappealing conclusion anyway.

According to the epistemic non-reductionist, intentional explanations are justified if they can be shown to have extra explanatory value. More in particular, the value of a intentional explanation is usually said to lie in that:

- they are able to grasp generalizations of kinds;
- they are, more in particular, able to grasp functional generalizations;
- they add predictivity, often due to the fact that we deal with generalizations or that we work with non-accidental properties of populations in stead of individuals.

Thus exactly *because* intentional explanations are different from physical explanations – instrumental, normative, rationalizing, generalizing and noisy – those kinds of explanations cannot be reduced to physical ones.

In chapter 7 I will discuss some applications of intentional explanations that indeed seem to have this extra non-reductive value. For now, I hope I have made a strong case for reading Dennett as a non-reductionist, and thus, to have shown that the intentional stance has to be read as more than merely a simple heuristics, but a type of explanation in its own right.

5 The Optimality Principle[43]

Up to this point, I have given a favorable and sympathetic account of Dennett's theory of the mind. In this chapter I will zoom in on Dennett's theory of design interpretation, understood as a part of his theory of the stances. Dennett's theory of design interpretation ("artifact hermeneutics") works according to the same logic as his theory of mind interpretation but it is also constrained by it. The result is what I call an optimality account of function interpretation. The main role of this chapter is to critically reconstruct Dennett's ideas on the role of optimality in technical function interpretations.

I have focused until now on Dennett's work on interpreting attitudes: the intentional stance. And for a good reason: the intentional stance is the irreplaceable heart of Dennett's theory of mind, and it has given him his important role in philosophical debates about the mind.

But the intentional stance cannot be properly understood without taking into consideration its other half: the design stance. The intentional stance may define Dennett's theory of mind and consciousness, but the design stance in an important way defines the intentional stance. It is the design stance that interests me in this thesis. Unfortunately, despite its crucial role, Dennett is terribly unclear about what exactly the design stance is, and he gives contradictory clues as to what *exactly* its role is supposed to be, especially when it comes to the interpretation of technical artifacts.

The design stance and the intentional stance are very closely related, and they share many characteristics. There is, however, a big question of what the hierarchy between the stances is. On the one hand, Dennett wants the design stance to play a heavy explanatory role in his theory of the intentional stance. Intentionality is then explained in terms of biological design. The design stance, from that perspective, explains the attributions we make from the intentional stance.

On the other hand, the very term *design* stance seems to require that we take into consideration the intentions of a designer. Often Dennett indeed suggests that taking the design stance simply means taking the intentional stance towards

[43] Parts of this chapter are derived from papers I wrote in collaboration with my colleague Krist Vaesen (Vaesen, K. and Amerongen, M. v.: forthcoming).

the creator of the design. The intentional stance, in that case, would explain the design stance. This sounds dangerously circular and Dennett is aware of that. Moreover, Dennett wants the design stance to cover both technical and biological design. And as a severe anti-creationist, it is quite odd to talk about intentions behind biological 'designs'.

To accommodate these issues, Dennett moves his theory of the design stance into the direction of what I call the optimality account. 'Design' in 'design stance' in such an account then means: design without a (real) designer.

Dennett's move towards optimality saves Dennett's favorite approach towards biological design. But it leads to a very awkward theory of technological design. It seems something must give...

5.1. The three stances

In *The Intentional Stance* (1987f), Dennett describes three basic types of stances we take towards entities in the world. From the *physical stance* we see objects in terms of their physical properties, roughly like physicists do (*explaining 'causes'*). Doing quantum mechanics relies on taking the physical stance, but predicting the trajectory of a ball in order to hit a home run does too. We could call the latter 'folk physics' (Dennett: 1991b), in order to distinguish it from it's more precise and systematic scientific counterpart.

Secondly, there is the – now familiar - *intentional stance*. When we use the intentional stance, we ascribe attitudes to the system at stake, that is to say, we interpret it an intentional system. We try to interpret the behavior of the system on the basis of the mental states we attribute to it, assuming that it behaves rationally (*interpreting 'reasons'*). As we have seen, we may use the intentional stance for a wide range of "agents": human beings, animals, artifacts, organizations, even inanimate objects.

Thirdly, there is the *design stance*. If we take the design stance, we understand an object as a designed object, and try to predict its workings (or try to derive its purpose) on the basis of the assumption that the design is optimal. When we take the design stance, we could say, we '*interpret the causes'*, inferring the reasons behind certain mechanisms or features of a designed item. As such, the design stance seems to be a hybrid stance, sharing features with both other stances. We use the design stance in two domains: function interpretation of technical artifacts ("artifact hermeneutics"), *and* function interpretation of biological functions (in, for instance, evolutionary biology). As we shall see later in the chapter, according to Dennett, we are dealing with design *without* a designer.

Dennett sometimes says that the design stance is exclusively meant to capture the designed *behavior* of systems (Dennett: 1987f, p. 16-17). This suggests, perhaps, that the design stance is only meant for (parts of) artifacts that have some 'autonomous' movement that we can predict, like a move made by a chess computer, or the sound of an alarm clock when it is time to wake up. But at other places Dennett applies the design stance more broadly, including artifacts that have no behavior in the ordinary sense of the word, for instance, hand axes and cherry pitters (see especially Dennett: 1990, p. 184).

It is true that the most general formulation of the design stance is that it helps us predict behavior of designed items on the assumption that the item will perform as it was designed[44] to perform, and on the assumption that it will not malfunction:

> "I simply *assume* that it has a particular design – the design we call an alarm clock – and that it will function properly, as designed. (...) Design-stance predictions are riskier than physical-stance predictions, because of the extra assumptions on board: that an entity *is* designed as I suppose it to be, and that it will operate according to that design – that is, it will not malfunction" (Dennett: 1996, 29)

Intuitively, the idea that the design stance somehow helps us predict certain kinds of behaviors of certain kinds of objects is very clear. 'Push the button and the ventilator will blow'. 'Pull the rope and the light will go on'. 'Press some buttons on the remote and the movie will play'. And clearly, we indeed apply the design stance regularly in such a way when interacting with technical objects and this makes our lives easier.

One could even seriously question whether life would be possible for human beings in a pervasively technical world, without being able to take some kind of stance that resembles the design stance (cf. Preston: 1998a, in which she forcefully argues that tool use is at least as characteristic and paradigmatic for human behavior as using a language).

[44] Note that this description of the design stance suggests that design requires a designer, but this is not the case, as we shall see later.

But the push-the-*x*-and-it-will-y application of the design stance, appealing as it is in its simplicity, has a too limited range, also for Dennett's purposes. The design stance as formulated above takes for granted that we know what a particular item was designed for, *what it is*, and it even assumes that we more or less know how it is designed. In the example of the alarm clock, we will not only have to know that it *is* an alarm clock, but also that alarm clocks can be set by pressing and turning buttons, that they must have batteries, that alarm time is represented in a different way than actual time, and so on. Usually we just *know* these things, by convention perhaps, or intuitive design. But what if convention fails to give us an answer to what a certain item is supposed to be? This is certainly the case when we encounter artifacts from the past, and when we are dealing with biological 'artifacts'.

Knowing what something is for and predicting its workings on the basis of that knowledge is one thing. Finding out what it is for is yet another. Knowing what something is for is obviously a necessary condition for taking the design stance towards an object in the general sense. Dennett recognizes that there must be more to interpreting design and has indeed also focused on what we may call 'investigative' questions of design, questions like 'What is the function of this thing?', 'Why is this item constructed like this', and 'What could this object be?' The reason for this will become obvious shortly, and is to be found especially in the fact that Dennett wants his theory of the stances to be applicable in (evolutionary) biological contexts too, explaining for example what a certain biological feature might be for. In such contexts, of course, knowledge about functions and reasons of designed features are typically sought instead of known or assumed.

5.2. Two of a kind

The physical stance ranges over all physical objects in the universe. Indeed, every event in the universe can in principle be explained and predicted from a physical point of view – the physical stance. The relation between the intentional stance and the physical stance has now been discussed at length. But how does the design stance fit in this scheme?

Dennett often treats the design stance as a third stance, but the design stance shares so many characteristics with the intentional stance, that we could make a more adequate distinction between the physical stance on the one hand, and the pair intentional stance – design stance on the other hand.

The design stance and the intentional stance share, in the first place, that they are in some cases more efficient than the physical stance. Like the intentional stance, the design stance is practical due to the fact that it takes an extra assumption on board and gains predictive leverage in doing so[45]. We have already seen that the intentional stance requires us to assume that the agent, whose actions we are trying to make sense of, is rational (the rationality assumption). We have to make a similar assumption when we take the design stance: we have to assume that the design is optimal. Take, for example, Dennett's favorite example of the design stance in action: reverse engineering[46]. Reverse-engineering is what engineers do when they want to fathom the design of an artifact of a competitor. It is an interpretative process in which the function of an artifact is reconstructed on the basis of the physical constitution of the artifact — what the artifact can do — *and* on the assumption that the design is optimal. The latter means that every component of the artifact under scrutiny has a *raison d'être*. We have to assume that the designers of the object did nothing in vain (reverse engineering will be discussed in more detail below).

The rationality and optimality assumptions limit the number of behaviors or functions that the system could have. A chess computer can make numerous moves, but only a limited amount of them are smart moves - and we can expect (and will assume) the computer to make smart moves. And a vacuum cleaner can perform an endless number of functions (as a place to store valuable goods in, as a paper weight, as art, as something to sit on, as material to make plastic cups of...), but it wouldn't be especially good in fulfilling these functions. There are better chairs than vacuum cleaners, and the capacity of the vacuum cleaner to suck up dirt would, if used as a chair, be superfluous – a waste.

The design stance and the intentional stance are essentially different from the physical stance. They are both "teleological" in their subject. That is, contrary

[45] One could argue that the physical stance itself relies on certain assumptions, for instance, the assumption that every event has a cause, that simple explanations are better explanations, and that we can describe the physical world in mathematical language. But these are different kinds of assumptions (or rather: rules) than the ones we adopt in the case of the intentional stance and the design stance because they do not rationalize the explanation.

[46] See for instance Dennett: 1995b, p. 212-213.

to the physical stance, they refer to purposes: goals in the case of agents, and functions in the case of artifacts.

All three stances are normative in a certain weak sense that they prescribe how we should explain the phenomena that fall under them. But the intentional stance and the design stance are normative in a particular sense. When we take the intentional stance or the design stance, we have certain normative expectations about the behavior of, respectively, agents and artifacts (cf. Hurley and Nudds: 2006).

We expect the behavior of agents to be rational and the design of an artifact to be optimal (or rationally designed). Stronger even, the expectation or assumption of rationality and optimality is constitutive (Child: 1994, 8-9): *constitutive* in the sense that their successful application just makes the attribution and characterization of the object at stake true (e.g. when we successfully ascribe a belief to an object, the object has that belief, and is an agent). For Dennett this even means, paradoxically, that there strictly are no agents but (pretty) rational agents and thus no designs but (pretty) optimal designs. We only 'see' agents if we view them as being rational, and we only 'see' design if we view artifacts as having some optimal function[47]. When we predict or explain a certain intentional action, we only take into consideration those actions that would be reasonably smart to perform, given the agent's cognitive capacities, its current environmental condition etc. Similarly, when we look at an artifact's function, we only take into consideration the reasonably smart designs that would allow the artifact to perform its hypothesized function. By contrast, when we look for the cause of an event in physical nature, we are not looking for 'smart' causes, or optimal causal relations. So, both the design stance and the intentional stance have a normative condition (i.e. rationality) that the physical stance lacks.

Let me repeat here shortly that given the differences between the intentional and the design stance on the one hand, and the physical stance on the other, most Dennett interpreters haven taken the theory of the stances to mean that *only* the physical stance describes the real world and that the other two stances are *merely* instrumental - practical, heuristic methodologies. That is to say: the items that are described from the physical stance exist independently from the

[47] This, of course, derives again from the principle of charity that is inherent in Dennett's interpretationism. For a critique on the necessity of rationalizing intentional descriptions and the contradictory results of such a position, see especially Stich: 1990.

predictive or explanatory strategies of epistemic agents. Design and intentional states, on the other hand, rely on normative assumptions *we* impose on the world - because we find it practical. Often it is therefore held that the constitutive nature of the intentional stance and the design stance gives their objects a separate ontological status. The status of physical states is rather clear: physical states exist, out there, in the real world and they are correctly or incorrectly described by physical laws. Not so for the status of beliefs, desires, functions and purposes. Rather than entities in the real world, out there, they are *ascribed* entities, that exist in virtue of us. As discussed in 3.6, I do not believe that the constitutive nature of the intentional stance and the design stance necessarily means that we should be ontological anti-realists or reductionists about teleology, but shall not further argue for that in this thesis. What matters is whether these stances are *epistemically* valuable, despite or due to their own nature.

Summarizing, the theory of the stances consists of three stances, the physical stance, the design stance, and the intentional stance, where the latter two are of the same kind: they both rely on normative assumptions that can be seen as constitutive for their referents. But this is still very general, especially with respect to the design stance. That the design stance relies on the assumption that we have, in some sense, to do with a 'optimal' design is clear enough, but what does optimality mean exactly and how do we determine an artifact's optimal design?

5.3. Four claims about the interpretation of artifacts

I have argued that the intentional stance and the design stance share many similarities. But their relatedness goes much deeper and this has consequences for and limits the way Dennett works out the idea of the design stance as what I will shortly refer to as 'the optimality account'. To explain that, I have formulated four theses that I believe capture Dennett's theory of artifact interpretation. Let me first sum them up[48]. I will explain all four of them in detail in the coming paragraphs.

[48] Thesis IA1 is a central thesis in Dennett: 1987f (but argued for in many other papers). Thesis IA2 is best developed in his Evolution, Error, and Intentionality (Dennett: 1987a). Theses IA3 and IA4 are directly derived from Dennett's paper on artifact hermeneutics (Dennett: 1990), in

(IA1) **thesis of attitude generosity***:* We (may) ascribe attitudes to technical artifacts, just as we (may) ascribe them to human beings and other biological organisms.

(IA2) **thesis of design primacy**: When in doubt, the (biological) design stance has explanatory primacy over the intentional stance.

(IA3) **thesis of function generality***:* Function ascriptions to technical artifacts are not interestingly different from other function ascriptions, such as to biological, or other functional items.

(IA4) **thesis of optimality**: the function of a certain item is – or should be understood as – what it is best able to do (or be), given its physical constitution and its context, and not what the designer(s) or user(s) intend it to do (or be). Function interpretation, thus, is not –or should not be- a matter of interpretation of intentions.

The second part of this chapter is dedicated to interpreting Dennett's ideas on design on the basis of these four theses. Let me first shortly introduce them.

The thesis of attitude generosity (IA1) forms the core of Dennett's theory of mind. We might also call it 'the thermostat thesis', after Dennett's claim that even the behavior of a thermostat can be legitimately be explained and predicted from the intentional stance. The principle of generosity (IA1) is by itself a very large and radical claim, about which one could write a whole dissertation. But the claim is easy to accommodate within the interpretationist framework that I assume to be correct in my dissertation (as amply discussed in chapter 2). That is to say, we will need an account of attitudes as one in which an attitude does not have to be conscious, or represented in the agent. And we will need to accept that there is a broad range of beliefs, ranging from proto-beliefs to full blown represented beliefs. Both are really just a matter of definition which has to show its worth, and they are both easy to accept for the interpretationist. In chapter 7 I find further support that a broad and gradual account of attitudes as Dennett

which he argues for a generic account of the interpretation of biological items, technical artifacts, as well as texts and people's mental states.

defends has shown its worth and that it complements the interpretationist view on the mind nicely.

Thesis IA1 is further justified in the second thesis, the thesis of design primacy (IA2). IA2 basically says that Dennett is a biological naturalist with respect to issues of the mind. IA2, therefore represents Dennett's idea that the mind is in the end a biological organ, that questions of mind are thus biological questions and that these questions are well posed in an evolutionary-biological context. Minds, then, are in the end explained by the (biological) design stance, and the intentional stance can thus be reduced to it.

Dennett's broad gradualist approach to attitudes (IA1) fits this evolutionary framework nicely. After all, once we accept that minds have evolved somewhere in evolutionary history, to the complex kinds of minds that human beings have, and that even the very simple ways of information gathering and simple drives can be seen as (proto) beliefs and (proto)desires (Dennett: 1996), it is only a very small step to adding that simple artifacts such as thermostats can be ascribed attitudes.

The thesis of function generality (IA3) is a consequence of Dennett's biological naturalism as well. As a biological naturalist, Dennett wants to understand culture in the same terms, or continuous with, nature (i.e. biology), so he needs a generic notion of function.: nature and culture have to be understood in line with each other. In addition, IA3 forms the justification of this evolutionary-biological context (justification of function talk in biology).

Thesis IA4, finally[49], tries to tell us somewhat more concretely how the design stance must be applied and what it means. IA4 therefore says that as a generic stance (IA3) it *has* to be conceptualized in a non-intentional way. This I call the optimality account of the design stance.

[49] I could have added two more theses, namely:
IA5) the interpretation of the meaning of a text, the content of an attitude, or the function of an artifact is always indeterminate
IA6) there is no original intentionality
But (IA5) is simply spelling out Dennett's interpretationism about meaning. As I will work from the assumption that this interpretationism is correct, it is not a claim for discussion in the next chapters. IA6 is the main argument for IA3, and will be dealt with when I discuss IA3. I have therefore chosen not to discuss it further as a separate claim.

Theses IA1, IA2 and IA3 are theses that Dennett has repeatedly and clearly subscribed to. Thesis IA4, by contrast, requires more interpretation on my part than the other three. Dennett is a moving target (cf. Dennett: 1987d, x) when it comes to the concretization of the design stance, so in order to be able to evaluate it I have to fix his position on it. IA4 as I define it is, I will argue, the most charitable interpretation of the design stance at work, respecting the other three theses that have to be considered fundamental to his theory. Note that theses IA1 and IA3 are very much alike. IA1 wants equal treatment of attitudes of humans, other biological organisms, and technology, and IA3 wants equal treatment of biological functions and technical functions.

5.4. (IA1) Thesis of attitude generosity

We (may) ascribe attitudes to technical artifacts, just as we (may) ascribe them to human beings and other biological organisms.

The first thesis is about attitude ascription to artifacts. Dennett sees no particular difficulties in ascribing attitudes to many kinds of artifacts, even very simple artifacts, like thermostats, like we do to human beings. I have given some examples of this in previous chapters. What to make of this kind of interpretation of artifacts?

To get this question right, I will first clarify some important terms, that are often not very well distinguished.

Having a ***theory of mind***, means having the capacity of an animal or person to represent itself or others as having intentional, content-bearing representational states. (Griffin and Baron-Cohen: 2002, 85). Many cognitive psychologists call this theory of mind '***the intentional stance***', but this can be slightly misleading as Dennett uses the term not only as an actual skill, but also to refer to a methodology. The intentional stance, according to Dennett, may be applied both to agents that have a theory of mind, and those that don't.

Intentional systems, or ***agents***, are what we apply the intentional stance successfully to[50]. It is important to see that not every intentional system is able to take the intentional stance, i.e. has a theory of mind. At least, this is the way Dennett, and many cognitive psychologists (chapter 6) and ethologists (chapter 7), conceptualize the notion of intentional system or agent.

There are different kinds of intentional systems, some more complex than others. Dennett usually distinguishes between first-order, second-order, third-order, and x-order intentional systems (Dennett: 1987b). Human beings, then, are probably the only known third-and-higher order intentional systems. Thermostats, as well as most animals, are at best first-order intentional systems. Some monkeys show second-order intentionality, perhaps even third.

More complex intentional systems, like human beings, require a more elaborate intentional stance than simple intentional systems. Simple intentional systems may well be explained by means of a 'simpler' intentional stance. So we do not only have a (gradual) scale of intentional systems, we also have different grades in complexity of intentional-stance-takers. Many animals have ways of distinguishing animate entities from non-animate entities, and will adapt their behavior on the basis of this knowledge. This is an important cognitive skill, as it helps the animal to be extra alert for possible prey, predators, and mates (cf. Dennett: 2006). As we will see in chapter 6, children are able at a very early age to predict the movements of simple goal-directed "creatures", but it takes them years before they are able to reason properly about the intentional states of fellow human beings.

Dennett is often not clear about what he regards as 'being able to take the intentional stance'. At some points, he argues that every creature that is able to discern animate beings from non-animate beings, is able to take the intentional stance (e.g. Dennett: 2006). But in other passages, he suggests that taking the intentional stance requires some conceptualization of attitudes like beliefs and desires (e.g. 1987b). When doubt may arise, I will distinguish the 'generous' ***intentional stance***' from the '***full blown' or 'strict' intentional stance***'. The generous intentional stance, then, refers to the ability to predict action-like behavior, without explicit reference to mental states. More on this in chapter 6.

[50] I leave it whether it is the case that (1) whenever there is an agent, we take the intentional stance towards the object, or (2) whenever we take the intentional stance towards something, it is an agent. Dennett defends (2), but at this point it does not matter which one is correct.

Now, if we want to ask about the interpretation of *actions* of artifacts, we may distinguish a number of questions:

(1) *do* we interpret the actions of (some) artifacts by means of the intentional stance? This is an empirical question, on which there is not much material available. Yes, of course we do interpret actions of some artifacts by means of the intentional stance, but it seems to be an overkill to use the full blown intentional stance for artifacts (see also chapter 6)

(2) Is there a *correct* way to interpret the actions of artifacts? This is a question about methodology which is for instance extremely relevant in the context of Artificial Intelligence research where one would certainly want to discuss levels of action complexity, kinds of actions, etc., of artificially intelligent robots.

(3) *Are* these artifacts intentional systems, or agents? This is an 'ontological' (in the Quinean sense) question. As pointed out, Dennett believes that they are, whenever we can fruitfully apply the intentional stance to them.

(4) Are technical artifacts themselves able to take the intentional stance? Can they, for instance, distinguish between animate and inanimate things? Can they predict the trajectory of an animate entity? An interesting question, that I will however *not* discuss in this thesis – whether or not they can is irrelevant to their being interpreted as agents.

With respect to the interpretation of actions of artifacts, the philosophical weight usually lies on the third question: whether artifacts can be granted an ontological respectable status as agents. But from Dennett's paradigm, the answer is rather straightforward: a useful application of the intentional stance to any object makes it an agent and grants it "ontological" respectability. So artifacts do not require a particular approach for Dennett; whenever you accept general instrumentalism about intentional states, you accept that artifacts can be considered agents. Because this point of view is inherent to Dennett's project, and because it is plausible for at least human beings and animals (as I shall further argue in chapter 7), I will accept this thesis as a given, and seek quarrel elsewhere.

5.5. (IA2) Thesis of design primacy

When in doubt, the (biological) design stance has explanatory primacy over the intentional stance.

Dennett wants to naturalize (in the sense of biological naturalism) intentions and explain the mind as a product of natural evolution. I call this the thesis of design primacy. Dennett argues that whenever the intentional stance is unable to give us an answer what the content of an attitude is, we (might) investigate the design of the entity.

The thesis of design primacy is perhaps best introduced by shortly sketching the domain over which Dennett's three stances range. As Figure 2 shows, the physical stance ranges over all *physical objects*. Within these physical objects, two subsets can be distinguished: designed items in general, and within that set designed *agents*.

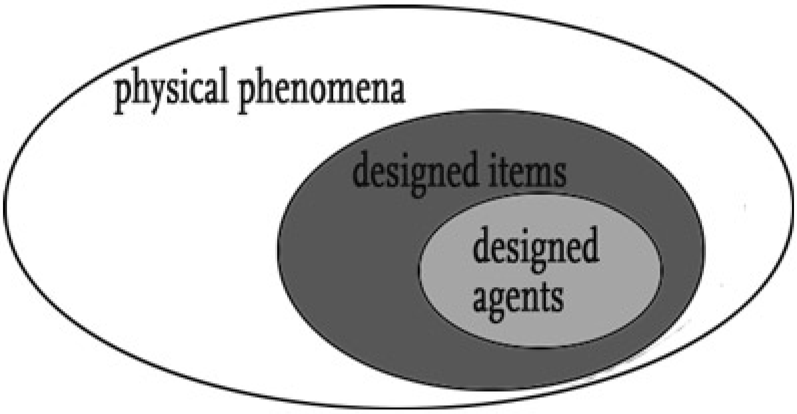


Figure 2: the referents of the three stances (physical stance for physical phenomena; design stance for designed items; and the intentional stance for designed agents)

The design stance ranges over all *designed* items (Dennett: 1987f, 16-17). Crudely stated: it covers those behaviors and features that have been designed to be there. The design stance thus spans a wide range of items, including mechanical clocks, typewriter bells, but also matches, chairs, computer programs, organs, eggs and trees. The intentional stance, in turn, covers a part of the domain of the design stance, i.e. designed *agents*.

We thus get a picture in which every agent is a *designed* agent (and every designed object obviously a physical object) and thus can be explained as being designed. The design stance is particularly useful for designed items in the broad sense and the intentional stance particularly useful for designed agents. But when the intentional stance fails to provide a satisfactory answer, we fall back on the design stance.

Take an artifact to which we ascribe certain attitudes, like a robot-vacuum cleaner (my example). Let's say we observe that this is a particularly smart vacuum cleaner; whenever a certain corner in the room is dusty, it moves towards it and starts vacuuming. Does it 'believe' (in the Dennettian sense) "that this corner of the room is dirty", or "that there is dust in that corner", or "that I haven't been in that corner for a while, let's go there, as there will probably be some dust over there"? In the last case, of course, the vacuum cleaner does not really have beliefs about dust at all. It is just "smart" enough to clean the spots that haven't been cleaned for a while, on the assumption that spots that haven't been clean for a while, will be dirty. (But also in this case, the intentional stance works.)

Now, if we want to know, for some reason or other, what the cleaner "really believes", it is probably wise to find out how the vacuum cleaner was designed. We could look at the mechanics of the cleaner, or ask the designers of the vacuum cleaner how it was made to perform. Note that interpreting design very quickly turns into interpreting designer intentions. Let's bracket that thought for a while and follow Dennett's reasoning.

Dennett claims that we can approach attitudes of animals, and other biological items, like plants and trees, *and* states of minds of human beings in the same way as we approached the smart vacuum cleaner. If we want to know what the exact content is of a cognitive or emotional state of an animal, we look at its evolutionary history, and try to find out how it has been designed. The answer to the question 'what does it think?', can be reduced to the question 'what is it designed to think?'

Does the frog believe that there is a fly coming towards it, or does it believe only that there is a small black something coming towards it? Or take the famous vertical symmetry detector that many animals are endowed with. The vertical symmetry detector helps animals distinguish other animals (as long as they are directly facing them at front): predators, prey, rivals and mates. What does an animal think when its symmetry detector detects a vertically symmetrical something? "A predator has spotted you", or "incoming food", or does it 'merely' believe that "Someone is looking at me!" Sometimes these questions can only be answered, Dennett says, if we look at the design specifications, and even then, we may not be able to come up with a satisfactory answer (see esp. Dennett: 1987a). States of mind are, for Dennett, just like every other part of organisms, designed features that are viable to design interpretation.

The thesis of design primacy applies to all attitudes, including those of human beings. This means that if you want to know more about the attitudes of a human being, you might get answers by looking at its (biological) design. This gets especially clear in those cases in which the intentional stance does not work – in cases of suboptimality, irrationality or psychological disorders. We may then have to drop the intentional stance, and look how we are built. "Design language" will quickly take the place of intentionalistic language. We will then soon be talking about dopamine levels and oral fixations, about impaired social modules, or problems in the affect structure, not about beliefs, desires, hopes and wishes.

In Dennett's words the thesis of design primacy reads:

> The migration from common-sense intentional explanations and predictions to more reliable design-stance explanations and predictions that is forced on us when we discover that our subjects are imperfectly rational is, independent of any such discovery, the proper direction for theory builders to take whenever possible. In the end, we want to be able to explain the intelligence of man, or beast, in terms of his design, and this in turn in terms of the natural selection of his design (Dennett: 1978c, 12)

5.5.1. Derived and original intentionality

Dennett's claim that the design stance is a final arbiter over the intentional stance has everything to do with the debate about derived and original intentionality. Artifacts, it is commonly agreed, have derived intentionality, whereas human beings have original intentionality. It is on the basis of this distinction that the analogy between applying the intentional stance to artifacts, and to human beings is usually rejected. For whatever attitudes artifacts might have, they are not 'original', but always derived from the intentionality of its creator.

Dennett rejects the distinction between original and derived intentionality on two grounds. First, because we can imagine artifacts having 'original' intentionality, secondly, because the intentionality of human beings -being artifacts of nature- is just as derived as that of artifacts. These two grounds serve to undermine the whole distinction between original and derived intentionality.

What would be an example of an artifact having original intentionality? Dennett gives an example: the survival-robot. He asks us to imagine that you decide that you want to keep your body alive over the coming 400 years, so that you would be able to see what the world will be like after that period. You decide to build a surviving capsule, which will wake you up at the appropriate time. (Dennett: 1987a, 295-296) As it is your life we are dealing with, the capsule better have enough energy to sustain through the ages. You add an energy plant close to the capsule, and figure that it might be a good to give the machine some mobility, so that it is able to look for its own energy just in case of emergency. We are still thinking about a machine that has to protect your life, so we will surely want that the machine can protect you as best as we can (science fictitious) think of. So, you add the ability to respond on changing circumstances to the robot, the ability to move to safer territory, to anticipate dangers and avoid them, etc. And we also have to take into account that other people will be inspired by your survival capsule (now already a robot), and build similar, possibly competing, ones. Etcetera and so forth.

> The result of this design project would be a robot capable of exhibiting self-control, since you must cede fine-grained real-time control to your artifact once you put yourself to sleep. As such it will be capable of deriving its own subsidiary goals from its assessment of its current state and the import of that state for its ultimate goal (which is to preserve you). These secondary goals may take it far afield on century-long projects, some of which may be ill advised, in spite of your best efforts. Your robot may embark on actions antithetical to your purposes, even suicidal, having been convinced by another robot, perhaps, to subordinate its own life mission to some other. But still, according to Fodor et al., this robot would have no original intentionality at all, but only the intentionality it derives from its artifactual role as your protector (Dennett: 1987a, 297)

Thus, for Dennett, this survival capsule, an artificial system that has goals of its own, would certainly be an ("original") intentional system.

Dennett's point is that we, human beings, are a lot like this survival robot. Drawing on Dawkin's (1976) idea of organisms being 'survival machines' designed to prolong the life of our genes), Dennett says:

> So our intentionality is derived from the intentionality of our 'selfish' genes! *They* are the unmeant movers, not us! (Dennett: 1987a, 298)

I will not deal with the problems that have been raised on this view. What I want to point out is that Dennett seeks to naturalize intentions (hence: the "principle of naturalism"), where 'naturalize' really means: biologize. The design stance (as being the naturalistic, biological stance) then 'decides' over the intentional stance, whenever we want to know more about the content or meaning of intentional states of either human beings, other biological organisms and artifacts. But, as we shall see in the next paragraph, Dennett can only hold on to this position under a specific interpretation the third thesis, the thesis of function generality.

5.6. (IA3) Thesis of function generality

Function ascriptions to technical artifacts are not interestingly different from other function ascriptions, such as to biological, or other functional items.

We ascribe functions not only to artifacts, but also to biological traits, social institutions, and many other items. We do this not only in daily life, but also in science, like evolutionary biology, economy and social science. Dennett likes to compare interpreting functions in nature with interpreting functions of technical artifacts. In fact, this effort to treat biology, artifactuality and humanity in the same terms is clearly a recurrent theme in his work. This should not be surprising. Biological naturalists are very reluctant to make distinctions between products of nature and products of culture, trying to understand both, eventually, in the same terms. For this reason, the *similarities* between taking the design stance towards artifacts, and taking the design stance towards biological traits are much more interesting for Dennett than the differences.

There are different ways to be similar, obviously. For my purposes it is enough to distinguish two: either biology is modeled on technology, or technology is modeled on biology. Dennett likes to model biology on technology:

> When we *adopt the intentional stance* towards a person, we use an assumption of rationality or cognitive/conative optimality to structure our interpretation, but when something goes wrong – when we find evidence of apparent sub-optimality or break-down (...) [w]e can no longer reside much faith in the agent's own opinions (...) What should we do? **Consult**

> **the designers** – just as we do with other artifacts. For we are artifacts, after all, designed by natural selection to provide reliable survival vehicles for our genes into the indefinite future (Dennett: 1990, 187, my bold)

According to Dennett, if we interpret natural functions we assume that the natural item at stake has been optimally designed (by "Mother Nature"), and try derive what nature 'must have meant' when she created it. In his earlier work on the intentional stance (Dennett: 1987d), Dennett used to speak freely in terms of 'adopting the intentional stance towards "Mother Nature",' clearly trying to show the analogy between interpreting functions of technical artifacts, and biological functions. From this point of view, natural selection (Mother Nature) figures as a designer in a similar way as a human being that carefully designed a complex artifact (Dennett: 1987a, Dennett: 1990).

> The chief beauty of the theory of natural selection is that it shows us how to eliminate this intelligent Artificer from our account of origins. And yet the process of natural selection is responsible for designs of great cunning. It is a bit outrageous to conceive of genes as clever designers (...) There is, I take it, no representation at all in the process of natural selection. And yet it certainly seems that we can give principled explanations of evolved design features that invoke, in effect, 'what Mother Nature had in mind' when the feature was designed (Dennett: 1987a, 299)

In evolutionary biology, Dennett's view of natural selection leads to 'adaptationism', or also 'Panglossianism[51]'.

> The strategy that unites intentional systems theory with this sort of theoretical exploration in evolutionary theory is the deliberate adoption of optimality models. Both tactics are aspects of *adaptationism*, the "programme based on the faith in the power of natural selection as an optimizing agent". (Dennett: 1987b, 260)

Adaptationism's recommended methodology for interpreting functions in nature is to first assume that every part or trait of the organism is there for a

[51] Pangloss was a character in a Voltaire novel, a caricature of the philosopher Leibniz, who thought that for everything there is a good reason, and that we live in the best of all possible worlds. The term 'Panglossian paradigm', from Gould and Lewontin: 1979, similarly refers to the adaptationist belief that natural evolution creates the best of all possible words.

reason. For instance, if we find a bird that lays four eggs, we first assume that this number is somehow optimal: two is not enough, five is too much (Dennett: 1990, 187-188, see also Dennett: 1995b. See Gould and Lewontin: 1979 for a critique of adaptationism). We then try to explain why four would have been the better strategy. Only if we cannot come up with a plausible story, may we grant that the trait is perhaps a bad adaptation, or maybe an "exaptation".

5.6.1. A non-intentional generic account of functions

As mentioned, Dennett often likes to model biology on technology, speaking freely of interpreting the intentions of Mother Nature. It then might seem that Dennett opts for an intentionalistic account of functions, in which we interpret the intentions of the designer. But these kinds of statements must be seen as metaphorical and provocative and must not be read in the wrong way: Dennett's generic account of functions is *not* intentionalistic. In some of his work, this remains rather ambiguous and confusing, so let me explain why Dennett's generic account should not be read as an intentionalistic account.

The most obvious reason that thesis (IA3) cannot be construed as an intentionalistic account is that biological functions simply are not *intentionally* designed. The theory of evolution is exactly meant to *counter* the idea that biological traits are intentionally designed – it is an anti-creationist theory. A Darwinist, like Dennett, must see the difference then between a biological function and an artifact function. Indeed, Dennett is one of America's severest critics of creationism, and related theories about intelligent design (see Dennett: 1995b; Dennett: 2006). So how does Dennett reconcile his generic account of function, with his anti-creationism?

In addition, there is a lurking circularity between IA2 and IA3 when construed as an intentionalistic account. If intentions have to be understood in terms of designs (IA2), it seems terribly circular to claim in addition that design has to be understood in terms of (designer) intentions.

It is illuminating to consider that Dennett spoke of taking the intentional stance towards Mother Nature especially in *The Intentional Stance* (Dennett: 1987d), a collection of papers that are completely dedicated to show the power of the intentional stance in a wide range of domains. The paper in which he defends using the intentional stance in cognitive ethology and evolutionary biology was discussed earlier at length in a special issue of the *Journal of Behav-*

ioral and Brain Sciences (Dennett: 1983). Dennett refined some of his ideas in his 'reflections' on some of the collected papers. In his *Reflections, Interpreting Monkeys, Theorists, and Genes* Dennett recognizes the problem of circularity in his treatment of the intentional stance in evolutionary biology:

> In the context of *BBS*, my coda defending the use of optimality assumptions by adaptationists and discussing the relationship of that tactic to the intentional stance seemed to be a digression, raising side issues that might better have been left for another occasion (Dennett: 1987g, 277).

For Dennett, we can treat natural selection as an agent ("Mother Nature"), but it turns out that an agent does not have to be a conscious person able to represent intentions – the design rationales do not have to be represented in order to do their work. Dennett often talks, in this context, about 'free floating rationales', rationales that steer the design process, without being literally in the mind of anyone (Dennett: 1995b, Dennett: 2006). We also find this move in Dennett's reflections, e.g.:

> Proper adaptationist thinking just *is* adopting a special version of the intentional stance in evolutionary thinking – uncovering the "free-floating rationales" of designs in nature (Dennett: 1987g, 277)

In addition, Dennett claims that his methodology is rather indifferent to the whether this rationale was actually represented by anyone:

> The difference between a design's having a free-floating (unrepresented) rationale in its ancestry and its having a represented rationale may well be indiscernible in the features of the design, but this uncertainty is independent of the confirmation of that rationale for that design (Dennett: 1987g, 286)

Using the intentional stance on invented, hypothetical agents? Isn't that too radical and at all plausible? I think not. Just remember that, according to Dennett, we can also take the intentional stance towards thermostats. The separate steps that constitute the process of natural selection are "blind", but what matters is that the process of natural selection in the long term *works* like a design process. It is a process in which smart moves are selected because stupid moves literally die out. That eyes evolved, and hearts, and that we can intelligibly ask what a certain biological trait is for, is not an accident. The evolutionary algorithm is an invitation to rationalize explanations of biological functions. It is for this reason that we can fruitfully take the intentional stance towards the process. The process of natural selection, on the long term, selects the better traits, and gets rid of the maladaptive ones, and can therefore be fruitfully treated as an agent.

Thus, taking the design stance towards a biological item means that we interpret its function in terms of the intentions of its designer – Mother Nature, just like we are tempted to do in the case of artifact functions. But this may not be taken *literally*. Taking the intentional stance towards Mother Nature is treating the process of natural selection as an agent in the metaphorical way, as a heuristic tool if you wish. We may also say that we *pretend* the biological world to be designed by some kind of ideal and rational designer.

Dennett's refinement of his ideas on interpreting functions, at least in biology, is confirmed in his later texts. In 2000, Ruth Millikan again confronted Dennett with his circular treatment of the design stance and the intentional stance (a similar argument is found in Ratcliffe: 2001). Millikan, a straightforward naturalist about intentions (Millikan: 1991), raises the issue of Dennett's free use of the intentional stance when interpreting design and interprets him as claiming that the intentional stance is more basic than the design stance:

> Dennett takes the intentional stance to be more basic than the design stance. Ultimately it is through the eyes of the intentional stance that both human and natural design are interpreted (Millikan: 2000, 55)

Millikan, in her paper, tries to convince Dennett to drop this position in favor of her own biological view of the mind *and* in favor of his own thesis of design primacy (IA2): that the design stance is eventually to be held prior to the intentional stance. Her own theory says that an intentional system is a *designed system* per se. Only those systems that have a proper design history (or, rather, those systems that display non-accidental rationality patterns) are candidates for intentionality. The assumption of a designer (of optimality) should be grounded, and should not be accidental (e.g. we have to have evidence that a process of natural selection has taken place).

> From enough apparent rational behavior one can infer design for rationality, just as one can infer design for seeing from good sight. And from design for rationality, one can infer real dispositions to rationality patterns, as opposed to mere temporary illusions of such dispositions. It thus ap-

> pears that the intentional stance must be underwritten by the design stance, rather than vice versa. Then, too, the fact that the organism is rational, indicating that the selection pressures have slowly designed it to be rational, serves as a genuine explanation of its behavioral patterns... (Millikan: 2000, 61-62)

In other words, that the intentional stance works on certain intentional systems is not an accident, but is due to the fact that these systems must have been designed by selection, in some way, to be viable to such intentional interpretations. Intentional systems simply *are* designed systems, Millikan claims. Millikan smartly plays on Dennett's naturalistic consciousness, of course, and she succeeds. In his response, Dennett admits that:

> I agree with her that the design stance is more basic, in the sense that she defends (...) Millikan is right in any case that it is no accident that the entities that succumb to the intentional stance *projectibly* must have been designed to do so (...) (Dennett: 2000, 342)

Dennett also recognizes that this has implications or his use of the intentional stance on 'Mother Nature':

> Use of the intentional stance in biology – the Mother Nature stance, you might say – is at least a convenient compactor of messy (and largely unknown) details into a useful interpretation label. It is *as if* Mother Nature had this or that "in mind" (Dennett: 2000, 342)

So this seems to be the correct interpretation: intentions are to be understood as the product of natural evolution, which can be understood as a design process without a 'real' designer, whose 'intentions' we can nevertheless interpret by using the intentional stance, but *only* in a metaphorical way. This means that:

> [a]daptationism and mentalism (intentional systems theory) are not *theories* in one traditional sense. They are stances or strategies that serve to organize data, explain interrelations, and generate questions to ask Nature. (Dennett: 1987b, 265)

Dennett started his theory of interpreting functions and rationales in nature by making an analogy with the interpretation of functions of technical artifacts. But his biological naturalism pushes him into the direction of a stance that is at best a heuristic device, aiming to detect *hypothetizised* intentions. But if Dennett wants to hold onto his claim of generality, this means that the interpretation of *technical* artifacts also should be seen in the non-literal, hypothetical way in which we assume a design to have been created by an idealized designer (Father Engineer, let's say), and refrain from any interpretations of 'real' or 'actual' intentions.

The consequence must then be that if Dennett wants a generic account, it must be one in which the role of intentional interpretation is on the one hand upgraded (for biology), and on the other downgraded (for technology). The result is what I call a general optimality account; a conceptualization of function as optimal function, applicable both to natural items and technical items. By contrast, conceptualizations of intentionalists emphasize the *differences* between artifact functions and biological functions by stressing the role of the intentions of the creator in the case of artifacts. This is not an option for Dennett and he makes that perfectly clear in his later work on the interpretation of artifacts. So let's take a look at this later work and discuss how this optimality account is to be construed.

5.7. (IA4) Thesis of optimality

The function of a certain item is – or should be understood as – what it is best able to do (or be), given its physical constitution and its context, and not what the designer(s) or user(s) in-tend it to do (or be). Function interpretation, thus, is not –or should not be- a matter of interpretation of intentions.

If Dennett wants to hold on to his thesis of generality (IA3), he will have to construe his account of technical design on the lines of his account of biological design. This means that he will need a notion of technical function that at best refers to 'hypothetical' intentions, and 'hypothetical' designers, and certainly not to 'real', represented intentions. In his most explicit paper on the interpretation of technical artifacts, *The Interpretation of People, Texts and other Artifacts* (1990), Dennett indeed makes it very clear that the interpretation of artifacts is better off without reference to intentions.

To be sure, Dennett nowhere really defines or clearly describes his account of artifact interpretation, rather tells us what it is *not*, so I will have to charitably construe it myself, basing myself on Dennett's sparse texts on the subject and modeling it on his ideas on the interpretation of biological features. I will call his account of artifact interpretation the optimality account of the design stance, to emphasize the importance of the optimality assumption and thus the analogy with the biological realm.

From Dennett's *The Interpretation of People, Texts and Other Artifacts,* Dennett frames his optimality account mainly as a negative account, i.e. as *non-intentionalism.* There should thus be as little reference to intention in our function ascriptions as possible.

Let me start with this *non-intentionalism.* Dennett rejects any account that puts designer intentions at the heart of the concept of artifact function, a position I call extreme intentionalism. I take Peter McLaughlin to be the clearest defender of such an extreme intentionalist account of artifacts. According to McLaughlin:

> An entity is an artifact and has a particular artifactual function if it is assembled, reassembled, or virtually reassembled with that particular purpose in mind (McLaughlin: 2001, 55)

McLaughlin's account has little constrains and leads to a very broad conception of what a function is. It is not necessary that the artifact *in fact* does what the designer *intended* it to do. It is also not necessary that we even *have* to perform a certain activity, in order to give a certain item a function. For instance, using a wooden stick in the woods for easy walking, is using it as a walking stick. And leaving a fallen tree as it is, for someone to use as a bridge, is giving the tree the function of a bridge. But it does mean that it conforms in some way to our desires, and requires at least some 'virtual' agency. McLaughlin cites Sorabji's Rule:

> Without at least a virtual artisan, there are no artifacts (...) Function conferring must involve some act of the will and the intellect, or a pro-attitude and a belief: that is, the function bearer must be considered to be in some at least minimal sense desirable or at least preferable to the available options (45)

The virtual effort must be realistically possible; the agent's approval must be in some way responsible for the fact that the desired effect takes place (46). But things that do not work properly, can still be artifacts with functions.

Thus, according to McLaughlin, it is the *intention* of the agent that primarily determines the function of an artifact, and not the material constitution of the artifact, the chance that the artifact will in fact do what the agent intended it to do, etc. McLaughlin's account is meant to show that the concept of artifact function cannot be used in biology – artifact function is too much tied up with human intentions and values.

Dennett explicitly rejects such extreme intentionalism:

> Consider how the intentional fallacy looks when applied to artifacts: the inventor is not the final arbiter of what an artifact is, or is for; the *users* decide that. The inventor is just another user, only circumstantially and defeasibly privileged in his knowledge of the functions and uses of his devise. If others can find better uses for it, his intentions, clearheaded or muddled, are of *mere historical interest* (Dennett: 1990, 186).

Designer intentions, according to Dennett, only indicate what the function the artifact was *supposed* to fulfill according to the designer. As such, theories of function that are based on designer intention are bad predictors. The optimality account is supposed to give better *predictions*, exactly because it does not refer to designer intentions. So, the theory at least is supposed to work as a *predictive tool* (I will get back to predictivity later):

> we can get *better* grounds for making reliable function attributions (functional attributions that are likely to continue to be valuable interpretation aids in the future) when we ignore (...) 'what the [designer] says' (Dennett: 1990, 194)

But Dennett goes further than that. He not only rejects intentionalism that puts *designer* intention at the core of the concept of artifact function, but he rejects any reference to intentions:

> what something is *really for now* is no more authoritatively fixed by the current user's "intentions" than by any other intentions (Dennett: 1990, 194).

In other words, neither the intentions of designers *nor* those of users determine the function of an artifact, or what the artifact is. Dennett's main arguments against intentionalism in general are that (1) intentions are unreliable indicators, because the content of an intention may be indeterminate and because there is no reliable way to specify the contents of an intention; (2) intentions are no relevant indicators of function. *If* intentions enter the picture somewhere, it is at the *end* of our interpretation, not at the start.

Dennett quotes Wimsatt and Beardsley:

> Judging a poem is like judging a pudding or a machine. One demands that it works. It is only because an artifact works that we infer the intention of an artificer (Dennett: 1990, 177)

We might alternatively call the optimality account the 'apparent intention' account: intentions turn up as *conclusions* of our reasoning, and are not the starting point. On the basis of a successful interpretation we might then conclude that "apparently they meant to use it for that purpose".

I will discuss and criticize Dennett's arguments against several intentionalistic accounts of technical functions in the coming paragraph (5.8). But let me first discuss what the alternative, positive account is. Unfortunately, Dennett is not very clear about it, but at least states that we should be looking for some "best" or "optimal" role the artifact could have. An intriguing but telling description of the optimality account is this one:

> it counts against the hypothesis of something *being* a cherry pitter, if it would have been a demonstrably inferior cherry-pitter. (Dennett: 1990, 184, my italics)

This is still a negative formulation, but if we may interpret it in a positive way it says that when we would find an artifact that would be perfectly able to, say, screw corks, it is a cork-screwer, no matter what the designer, or its users, intended it to be. So the best way to determine what an artifact *is*[52], is to look at what the artifact would be best able to do.

Dennett's prime example of artifact hermeneutics, reverse engineering, gives some further clues about what the optimality reasoning amounts to:

> Reverse engineering (RE) is the process of discovering the technological principles of a device or object or system through an abductive analysis of its structure, function and operation (...) It often involves taking some-

[52] Now, it is not exactly clear how Dennett wants us to read the optimality account. Is it a general theory of function (this artifact can be/has been used for X), or more specifically, (also) a theory of artifact conceptualization (this artifact is an A). I have interpreted it as a theory of conceptualization, which, for the Dennettian, boils down to a theory of function ascription (the function of an artifact determines what it is).

> thing (e.g. a mechanical device, an electronic component, a software program) apart and analyzing its workings in detail, usually to try to make a new device or program that does the same thing without copying anything from the original.[53]

Reverse engineering is indeed a illuminating example of a case where designer intentions do not matter. The goal is to reconstruct the artifact, no matter what the actual designers thought and wanted. Most descriptions of reverse engineering on the internet (mainly about reverse engineering software code) suggest that reverse engineering can be perfectly done without any reference to intentions. It is a mere description of 'Cummins-functions'. A Cummins-function is a minimal and non-normative notion of function and is described simply as a capability of a part of a system to contribute to that system (Cummins: 1975). But this can clearly not be the kind of reverse engineering Dennett refers to. In that case, he should be prepared to drop any reference to 'optimality' as well, for Cummins-functions do not combine well with normative notions such as optimality. Dennett's account, we may say, adds to that notion of function the idea that of all possible contributions a part may have for the larger system, it is the 'best' contribution that counts.

And indeed, if we look at Dennett's reference to reverse engineering again, it is clear enough that Dennett wants to go much further than a simple description of Cummins functions:

> **Why** did GE make these wires so heavy? **What** are these extra ROM registers **for**? **Is this a** double layer of insulation, and, if so, **why** did they bother with it? Notice that the reigning assumption is that all these "why" questions have answers. Everything has a *raison d'être*; GE did nothing in vain. (...) [T]his default assumption of optimality is too strong; sometimes engineers put stupid, pointless things in their designs, sometimes they forget to remove things that no longer have a function, sometimes they overlook retrospectively obvious shortcuts. Still, optimality must be the default assumption; if the reverse engineers can't assume that there is a **good rationale** for the features they observe, they can't even begin their analysis (Dennett: 1995b, 212-213)

[53] http://en.wikipedia.org/wiki/Reverse_engineering

We find the phrasing 'good rationale' again in this quotation. This is only another confirmation that Dennett wants to construct a unified account of design interpretation, where the hypothesis of some kind of ideal, hypothetical designer figures as an aid in our reasoning. The why-questions and what-for questions so typical for artifact hermeneutics, can also be seen perfectly in this light.

The optimality account thus has two prime elements: (1) *non-intentionalism* and (2) *normativity*: we look for what the item is '*best*' able to do. 'Being best able to do' can mean different things. For example, is it about what the artifact is best able to do right now, ever, or in some ideal state? Could an object have several functions? How do we determine what a certain item is best able to do? Would this be an objective fact? Is it an internal fact about the object, irrespective of (intentional) context? Is that even possible?

As mentioned, Dennett is not very explicit about it, but I think we can discern at least two different reasoning strategies for the optimalitist[54]. In the first case, we find a certain object, and consider what it could be, given the assumption that it must be optimally designed (design → apparent purpose). Let me call this type of reasoning bottom-up optimality reasoning. In the second case, we start with a given purpose and infer how this purpose might be best brought about (given purpose → design). Let me call this reasoning top-down reasoning.

In case of top-down optimality reasoning, we assume that the best possible solution has been found given a certain function or design problem. In the case of bottom-up optimality reasoning, we assume the object at stake to fulfill that function in the best possible way (under the circumstances), but there might be objects that are better able to fulfill that function.

Because Dennett cannot start with intentions (or, for that matter, purposes), he must be on the first line (bottom-up), rather than the second (top-down). And this seems right. Consider Dennett's example of an old computer mainframe that has become obsolete and is now used as an anchor. Dennett argues that the object is better seen as an anchor, than as a computer:

> ... a Dec-10 mainframe computer today makes a nifty heavy-duty anchor for a large boat mooring. No artifact is immune to such appropriation, and

[54] With thanks to Palmyre Oomen for drawing attention to this distinction and helping me develop it.

> however clearly its *original* purpose may be read from its current form, its new purpose may be related to that original purpose by mere historic accident – the fellow who owned the obsolete mainframe needed an anchor badly, and opportunistically pressed it into service. (Dennett: 1990, 184)

This is a rather clear example of bottom-up reasoning. We find an item, in this case an old computer, in a certain context, and infer what it would be best able to do given the circumstances. Top-down optimality reasoning would probably have given different results. For if we would have wanted an anchor, we would have designed a rather different object: something heavier probably, with a hook, and without the superfluous reset-button. There are far better anchors than this computer (top-down), but in the current conditions, the best possible function for the object is to figure as an anchor (bottom up).

The two types are, however, not as easy to distinguish as suggested above. *Are there*, indeed, better anchors than the one used in these circumstances? If someone needed something to anchor a boat with, and by lack of something better at hand, wasn't this the best possible anchor he might have used? Optimality need not go in the direction of perfection, but is rather a compromise between a purpose and the circumstances in which it should be realized. Dennett nicely points out:

> The customary disclaimer in the literature is that Mother Nature is not an optimizer but a "satisficer[55]" (Simon, 1957), a settler for the near-at-hand *better*, the good enough, not a sticker for the *best*. And while this is always a point worth making, we should remind ourselves of the old Panglossian joke: the optimist says this is the best of all possible worlds; the pessimist sighs and agrees. (Dennett: 1987b, 264)

The two types of reasoning, thus, are in a sense different sides of the same coin. But they *are* two different types of reasoning, the top-down approach reasoning from a rational, or pretty rational, designer, and the bottom-up approach reasoning from the design itself. Dennett's examples (the anchor, adaptationism's methodology), as well his claim that we should reason from optimality to intent, suggests that he sees his optimality as following the bottom-up approach. I will

[55] 'Satisficing' is a term from economics, imported into the philosophy of practical reasoning by Herbert Simon. It refers to a minimal form of rationality (adequacy) that seems easier to embrace than perfect rationality or optimality.

argue later it might be wise for him to embrace both kinds of reasoning, as they can complement each other rather nicely (7.6.1, 8.4).

5.8. Two arguments against intentionalism

Now that I have sketched Dennett's optimality account, I want to spend some time considering Dennett's arguments against intentionalist accounts of (artifact) function. Obviously, Dennett has his own reasons for *wanting* the optimality account to work, as I have just discussed at length, but he still has to make plausible that the optimality account is an adequate theory of (artifact) functions.

The plausibility of Dennett's argument for the optimality account relies heavily on his arguments against intentionalistic accounts. The optimality account is primarily defended on the basis of the claim that an intentional account of artifact function is unsatisfactory, but I believe he confuses a number of forms of intentionalism that should be well distinguished. On the basis of a critical interpretation of Dennett's argument against intentionalism, I will argue in the coming paragraphs that it fails. This means that I will now switch from a rather charitable interpretation of Dennett's ideas, to a *critical* interpretation.

Dennett's main argumentative strategy for the optimality account is to show that it is our *next best thing*. Take it, or leave it – if you want to be theorizing about artifact functions at all, you should be talking about optimal functions, not intended functions. The optimality account thus is foremost a *negative* account that is supposed to construct a notion of function without a notion of intention. In fact, Dennett brings up few arguments that speak *in favor* of the optimality account and they are all weak (see 5.9). Let's first look in more detail at Dennett's argument against intentionalism.

The notion of artifact function is often defined in terms of intention, both in our ordinary understanding of artifacts (cf. chapter 6), as well as in philosophy. A popular philosophical account is the original intended design account or designer intention function account (exemplified by McLaughlin: 2001 and introduced on page 106 as extreme intentionalism) in which the function of an artifact is defined (primarily) in terms of the intention the designer had in mind when he created the artifact. Dennett's first attack on intentionalism is against such theories that reduce the function of an artifact to designer intentions.

In his *The interpretation of people, texts and other artifacts,* (Dennett: 1990) Dennett is inspired by a variant of "the intentional fallacy". This is a term from literary criticism, used by anti-intentionalists, that limits the meaning of a text strictly to the actual text. It is a direct response to intentionalists who think the meaning of a text can and should be reduced to, or related to, author intentions. The debate over the intentional fallacy started in 1954 when William Wimsatt and Monroe Beardsley wrote their article *The Intentional Fallacy* (Wimsatt and Beardsley: 1946), and continues still. Wimsatt and Beardsley argue against the interpretation of author intentions as being relevant for interpreting an art work. The debate has crystallized around a number of battle-points: (1) intentions of the author are often unavailable and our methods of tracing them are unreliable, and (2) intentions of authors are irrelevant to our interpretation of their creations. Dennett follows the line of thought exemplified by Wimsatt Beardsley, and extends their line of reasoning to artifacts (a text is an artifact in at least some sense, after all).

5.8.1. The indeterminacy argument against intentionalism

Dennett claims that designer intentions are better ignored when we interpret the function of an artifact. His main argument is that intentional attributions are unreliable because they may be wrong and because they are indeterminate (Dennett: 1990, 180).

Dennett uses mainly general examples where interpretation of author intentions is indeterminate or unreliable, and just claims that interpretation of designer intentions befall the same problems. And this seems right. Why should the interpretation of the intentions of, say, the author of a text, be harder or easier, or more or less reliable, than the interpretation of the intentions of a designer of an artifact?

Indeed, we can easily imagine many cases in which it is hard to determine what the intentions of the designer of an artifact were. Often we just do not know what the intentions of the designer were, for instance in the case of older artifacts. But also in those cases where the designer is known and still lives, it may be that he is unable to say exactly what he intended. It is therefore often hard or even impossible to get a reliable specification of the content of an intention, both ‘from the inside’ and ‘from the outside’. So, parallel to the interpretation of texts, specification of the intentions of the designer may be impossible or ambiguous. As such, designer intention is an epistemically weak starting point for ascribing functions to artifacts.

Dennett suggests that this is sufficient reason to reject an intentionalist account of artifact function, more specifically, an original intention (designer intention) account of artifact function. He reasons as follows:

(1) Interpretations of designer intentions are indeterminate and (therefore) unreliable.

(∴C) Interpretations of designer intentions are better ignored when we ascribe functions to artifacts

Let us suppose that (1) is indeed true. Then (C) does not follow. (C) only follows if we add at least a second premise, and I see no viable candidate for this second claim for Dennett to accept.

The conclusion (C) only follows with the premise that *indeterminate* function attributions are *bad* function attributions that we should avoid (or some such assertion). Dennett needs to defend *at least* some thesis like:

(2) Interpretations of intentions should be reliable and determinate in order to enter our ascriptions; so, (designer) intentions should be reliable and/or determinate in order to enter our functional ascriptions to artifacts.

But Dennett is well advised not to accept (2). After all, if we take the intentional stance, our ascriptions may also be indeterminate and unreliable, but this does not mean, as Dennett himself has repeatedly emphasized, that we should refrain from taking the intentional stance towards agents. On the contrary, we put the intentional stance to good use in daily life, and in science. Taking the intentional stance is justified whenever we can put it to good use, and it often is, *despite* the indeterminacy of its content. Dennett, then, should show that functional ascriptions that refer to intentions are *too unreliable to be of good use* – that is his own standard. Were Dennett to accept (2), this would immediately hold for his theory of the intentional stance, and hence undermine his whole theory of attitude ascription.

What follows from Dennett's unreliability argument at best is that the interpretation of intentions of the designer (or any other agent) is a very difficult task.

Sometimes there may be several, conflicting, hypotheses about what the designer intended. It may even be so that the artifact clearly suggest that the designer had a certain intention, whereas the designer himself says that this intention never existed. Furthermore, intentions may change during the process of creation, or may be different to regain from hindsight. Some artists even claim that the process of creation is not an intentional process at all; the same may very well be true for artifact design.

But that we sometimes may be epistemically unable to trace author or designer intentions does not mean that we should refrain from even trying. If tracing intentions is *sometimes* impossible, it could simply mean that asking for the thoughts of a designer behind an artifact is a legitimate question, but one that cannot always be answered. And if it *is in principle* impossible, or unreliable, to speculate about designer intentions, it would probably be better to conclude that we should not even begin to try to interpret art works or artifacts and stop talking about functions altogether. This would be the "behaviorist" approach. So epistemic uncertainty about intentions does not speak against an intentional notion of function per se *and* specifically: *if* they matter we have to take them into consideration, and if Dennett believes that this would lead to dubious science, he should become a behaviorists. Dennett does not want to be a behaviorist, so he shall have to show that they do not matter.

If I am right, Dennett's claim against the interpretation of design by means of designer intentions is not supported by the indeterminacy claim, and he should therefore find another argument to support his claim (IA4) that (designer) intentions do not matter for function ascription. Otherwise, the optimality account of function, as defended by Dennett, cannot be presented as *the best next thing*. An intentional account of function is still in the race, and should be weighed against Dennett's alternative, the optimality account.

5.8.2. The irrelevancy argument against intentionalism

Analogous to the debate about the intentional fallacy, Dennett makes a what I will call irrelevancy claim. He claims not only that designer intentions are unreliable (the indeterminacy claim from 5.8.1), but also that they do not matter (enough) to enter our interpretations. (A weaker version may say that our interpretation should at least not be *centered around* designer intention).

In the debate in literary theory about the intentional fallacy, the irrelevance claim says that the interpretation of art is about art, not about author intentions. William Beardsley, for instance, in a work on the intentional fallacy, argues that it is hopeless to derive the meaning of the text on the basis of the intentions of the author because (1) some texts have no authors (think of a random text generator) (2) meaning can change after the authors death, but his intentions cannot (3) a text can have meanings an author is not aware of (Beardsley: 1970, 18-19). In this line of thought a text does not derive its meaning from the intentions of the author – meaning is in the text itself. Most commentators agree that we should be interested in understanding the work itself, but they disagree about the question whether we have to know something about the intentions of the author, in order to understand the meaning of the text.

This point has, of course, been the subject of much controversy, also in the philosophy of language. Is it utterance meaning of the author we are after (as the intentionalist wants it), is meaning contained in the text itself (as the non-intentionalist will claim), or is meaning in the eyes of the readers (a different form of intentionalism)? *Extreme intentionalism*, a position nobody wishes to defend in this domain, claims "that the meaning of an artwork is whatever the author intends it to mean" (Carroll: 1999, 75).

This is a good place to take a short detour, and distinguish between extreme intentionalism and intentional realism. I will need this distinction at several points later in the thesis. Extreme intentionalism reduces something (e.g. meaning, function) exclusively to intentions, usually designer intentions. Intentional realism is rather a claim about what intentions *are*, specified as being literal, concretely existing states of minds in the heads of people (cf. 3.6). An extreme intentionalist may be a realist about intentions, but not necessarily so (you might hold a theory of ascription and define, e.g., meaning in terms of ascribed designer intentions). And an intentional realist *might* favor extreme intentionalism for his theory of meaning, but certainly not necessary so.

Extreme intentionalism leads to what is called Humpty-Dumptyism: by changing his intentions, the author could give every sentence, every phrase, every word, the meaning that he wants (Carroll: 1999). This is highly implausible, not only for works of art, but also for artifacts. If I write a text, intending it to be funny, but miserably failing, the text is not going to be funny, no matter how hard I intended it to be. Similarly, my intending a chair to be an airplane, will not make the chair fly (i.e. make it an airplane, or an artifact that has the function of flying). Extreme intentionalism is rarely seriously defended. Yet, most agree that intentions enter the meaning of the text *somewhere*.

Dennett similarly claims that designer intentions do not fix the function of an artifact. No artifact is immune to losing its original function, and therefore we cannot infer from intended function to current function (recall Dennett's example of the computer that turned into an anchor).

Here I object. The irrelevancy argument is very strong against original (designer) intention accounts of functions, but it is *only* an argument against *such* intentionalist accounts. And even then, it only works against extreme versions of such original intention accounts of functions. Consider, again, the interpretation of texts. There are hardly defenders of extreme intentionalism (Humpty-Dumptyism) anymore. But there are still many intentionalists that do want to defend a more elaborate variant of intentionalism. For instance: 'modest actual intentionalism' and 'hypothetical intentionalism'.

Modest actual intentionalism holds that the actual intentions of the author do determine the meaning of a text, *as long as* they are supported by the artwork. Moreover, when the text supports two equally plausible hypotheses about the meaning, the intentions of the author decide. Carrol, proponent of modest actual intentionalism, says:

> Attributions of meaning, according to the modest actual intentionalist, must be constrained not only by what possible senses the text can support (...) but also by our best information about the actual intended meaning of the utterer and author in question. (...) For the modest actual intentionalist, the author's intention here must square with what he has written, but if it squares with what he has written, then the author's intention is authoritative. (Carrol: 1999, 76)

And if even this is too strong, one might opt for hypothetical intentionalism that:

> ...maintains that the correct interpretation or meaning of an artwork is constrained not by the actual intentions of the authors, but by the best hypotheses available about what they intended. (...) The meaning of a text is what an ideal reader, fully informed about the cultural background of the text, the oeuvre of the author, the publicly available information about the text and the author, and the text itself, would hypothesize the intended meaning of the text to be. (..) That is, the hypothetical intentionalist claims that the meaning of the text correlates with the hypothesized intention, not the real intention, of the author (Carroll: 1999, 78).

Hypothetical intentionalism, or perhaps even modest intentionalism, albeit both intentionalist positions, do take into account that author intentions can be overruled, but still give a considerable role to author intentions in the interpretation of art works.

Applied to artifact hermeneutics, we can easily imagine a form of intentionalism that says that designer intentions can be overruled in certain circumstances. Hypothethical intentionalism, though rejected by Carrol as not having much to offer over and beyond modest actual intentionalism, even seems like a position Dennett might want to experiment with! So Dennett has rejected at best an extreme version of the original intended design account, not possible (and more plausible) milder versions of it. Moreover, Dennett still hasn't rejected a user-intention account of function.

Wrapping up my argument against Dennett's argument against intentionalism, it is helpful to distinguish several forms of intentionalism:

(1) O*riginal designer intention function*. Extreme intentionalism is the clearest example of this position that says that the function of the artifact is to be determined on the basis of what the designer intended the artifact to be. Dennett argues against extreme intentionalism, but neglects modest forms such as modest actual intentionalism and hypothethical intentionalism.

(2) *Intentional realism*. Strictly speaking, intentional realism is not even a claim about artifact interpretation at all, but a theory of what attitudes (including intentions) are. Dennett's indeterminacy argument against intentionalism is mainly directed against intentions as understood by the intentional realist, not against other theories about intentions, including his own interpretationism.

(3) "*user function*", i.e. the function for which someone or a group intentionally uses the artifact. Dennett shortly dismisses this position with the indeterminacy argument, but this argument only works against intentional realist accounts of intentions.

(4) *strict or strong intentionalism*. Mentioned earlier in 5.4, I want to distinguish strict intentionalism from generous intentionalism, because I will need the distinction later. Strict intentionalism refers to the full blown state into which minds can evolve or develop (like in most human beings). It is opposed to generous intentionalism that admits the

> existence of simpler, 'unconscious', minds, Applied to artifact interpretation, the strict intentionalist would claim that only designers (or users) with 'strict' minds can create an artifact, whereas the generous intentionalist might grand simple minds this capacity as well. I call the position that says or presupposes that strict intentions are necessary for understanding artifacts 'strong intentionalism'.

Concluding, Dennett defends his own optimality account primarily on the basis of his rejection of intentionalism. Dennett rejects intentionalistic accounts because reference to intention is unreliable and indeterminate (5.8.1). But unreliability and indeterminacy are inherent in our ascriptions of *any* state of mind, at least so for Dennett. And the irrelevancy argument (5.8.2) at best argues against a straw man position, Humpty-Dumptyism, that hardly anyone defends anymore, and ignores possible alternatives. And worse: the unreliability argument even undermines Dennett's whole theory of the intentional stance, because it relies on the claim that explanations may not be determinate – a claim that Dennett has worked so hard to *reject*!

5.9. Arguments for the optimality account

Dennett presents his optimality account as an alternative for the failing intentionalist account. But intentionalism has not been rejected yet. What could be positive arguments for the optimality account? I discern three directions:

(1) The optimality account fits Dennett's attempt to create a unificatory account of functions

A notion of function devoid of intentionality is of course preferred by Dennett, because he wants a notion of function that can be equally applied in both technology and biology. A unificatory account of course fits Dennett's thesis of function generality and his attempt to construct a unionist philosophy. But that Dennett *prefers* a generic account is a good reason for him to defend it, but not an *argument* for the optimality account.

(2) The optimality account gives a better description

Dennett may claim (and should show!) that when we ascribe functions to artifacts, we do in fact work according to the optimality strategy. In the next chapter I will criticize this descriptive claim about functional reasoning. Our daily concept of an artifact is still "generously" constructed in terms of intended design: ask a person what a certain artifact really is, and he will probably try to frame it in terms of intended original design. In such everyday function attributions, we rely heavily on the notion of intended design, both of natural items, and of technological items.

(3) The optimality account gives better predictions

The third argument for the optimality account is that it gives better predictions than any alternative. To be sure, Dennett *claims* that the optimality account has predictive force (see the quotation on page 107), but he never indicates why this should be so. What is there to predict? I will argue later (7.6.1) that the underlying thought could be that our hypotheses about what the function of an artifact is help us predict what it will do. For instance, on the basis of the hypothesis that something is a gun, we might predict that pulling the trigger in the right way will have it fire a bullet. Such predictions may even help the interpretation, for if the object does not fire a bullet, we might have to conclude that our initial hypothesis was wrong (this is prediction in its classical role of falsification, which is a good analogy also because falsification requires a specific as possible prediction).

I believe that predictivity is not going to settle the issue between intentional accounts of artifact function and the optimality account. I shall argue for this in chapter 7, where I discuss the methodological utility (amongst which predictivity as a methodological virtue) of the intentional stance and the design stance.

The general conclusion of this chapter is that Dennett does not provide many positive reasons for embracing the optimality account, and that his negative arguments fail to hit the right target. The optimality account fits his theory of artifact interpretation all right (theses IA1, IA2, IA3), but the optimality account would clearly be too weak if it were only supported by a reasoning like 'this is the only account that fits the rest of my ideas'. If I am correct that IA4, the optimality account, is indeed the theory of artifact interpretation that Dennett wants to defend, further support is needed. In the next two chapters I examine whether further support for the optimality account can be had by measuring the optimality account on the two standards empirical correctness and methodological value.

6 Taking a Stance as Daily Practice[56]

I have claimed that the optimality principle could be enforced if it had empirical support. If human beings do in fact interpret technical artifacts on the basis of optimality reasoning (rather than intentional reasoning), this would certainly count in favor of the optimality account, as it would then live up to the empirical standard. In this chapter I discuss some recent cognitive psychological investigations of humans interpreting technical artifacts in order to see whether the principle indeed finds such support.

Dennett's theory of the stances is a theory about how we (should) perceive and approach certain objects: physical objects, designed objects and agents. Cognitive psychologists have been inspired by Dennett's work on the stances (even literally using the same terminology) and have studied these ways of looking at kinds of objects empirically.

Before I start I should note that although the cognitive psychological researchers whose work I will discuss adopt the Dennettian stance-terminology and explicitly refer to Dennett, they turn out to work out the term 'design stance' in an importantly different way than Dennett. That is to say, the research does, like Dennett, address the way human beings approach designed objects, but it gives a rather different content to what it means to interpret design than Dennett does. Furthermore, the cognitive psychologists under discussion speak in terms of conceptualization of artifacts rather than in terms of ascription or interpretation. In this chapter, I will address the cognitive psychological construction of the design stance. I will use the term design stance very generally as 'the way we think about artifacts', where 'thinking about' means 'conceptualizing' in the cognitive psychological framework. In those instances where I restrictively use Dennett's original notion of it, I will explicitly mention it.

[56] Parts of this chapter, those about the design stance, are derived from a conference paper I wrote together with Giacomo Romano (2004). I thank Giacomo for his continuing insistence to bring this empirical material on the (development of the) stances into my thesis.

The empirical research by cognitive psychologists provides a wealth of data on all the three stances. Some research is done with adult subjects, but most of it is developmental. Both give us insight in our 'normal' way of looking at these kinds of objects and thus our 'normal' way of taking the stances. It is exactly this normal way of looking that I am interested in.

Strictly, developmental data only gives us insight in the *direction* of our way of looking at things. Applied to the conceptualization of function: if it were shown that most human beings, now and in the past, in the West and in the East, develop an intention-based concept of function, this is good reason to think that intention is (normally, typically, usually) at the core of the concept. In comparison, that functions are nowadays not ascribed to biological items might be a rather accidental property of the concept of function – accidental to the relatively recent theory of natural selection, and accidental to what children learn in school. Of course, there are no straightforward boundaries to be given (does 80 percent have to agree with the definition? 90 percent?[57]). Nevertheless, such research gives a rather good idea of how we generally perceive physical objects, designed objects and agents and is as such a proper[58] frame of reference to test the empirical correctness of the theory of the stances.

6.1. The stances in cognitive psychological research

The emphasis of the cognitive psychological research on the stances lies on our understanding of physical objects (causality) and on the ability to explain the actions of human agents in terms of their mental states: the Theory of Mind or what I have called the full blown or strict intentional stance – as opposed to the generous intentional stance (5.4). What interests me most, however, is the recent research that has been done on the generous intentional stance and the design stance.

Let me shortly say something more about this Theory of Mind. Human beings develop a theory of mind approximately at the age of four, that is, when they pass the false belief test. The false-belief test has several variants, but its main point is to test critically whether subjects are able to reason in terms of

[57] 'Normality', in contexts like these, is a hard to define term. A Millikaneske notion of normality usually works rather well in these contexts (see, e.g., Millikan: 1991)

[58] In fact, I believe it is the best way we *have*.

other person's mental states. For instance, subjects, usually children, are shown how the experimenter puts candy in one of two boxes. A second person that is present sees this happening (and the subject sees the second person seeing this happening), and then leaves the room. The experimenter removes the candy from the box, and puts it in the other box. Subjects are then asked where the second person will look for the candy on his return. Children younger than four years old will generally answer that the second person will look in the correct box (not realizing that the second person does not know that the candy has been moved). Older children will correctly say – pass the false belief test - that the second person will look in the original box (Wimmer and Perner: 1983, Wellman: 1990, See Griffin and Baron-Cohen: 2002 for a good overview, see also 5.4.).

Having a theory of mind is an important phase in the socio-cognitive development of a human being. Cognitive psychologists are beginning to understand more and more how this ability develops gradually over the first four years of a child's life. Results now point at the conclusion that children are able to ascribe *purposes* ("emotional states") to animate entities at a very early age. It is their reasoning about *belief states* that lags, exactly because beliefs can be *false* (see Griffin and Baron-Cohen: 2002, 91). Current research suggests that the Theory of Mind (the full blown intentional stance) develops out of a 'naïve theory of rational action' (which is very comparable to what I have called the generous intentional stance), a stance that shows little sensitivity to agent *beliefs*, but is sensitive to *purposes*. The naïve theory of rational action is an incomplete theory of mind, because it will generally fail to predict actions that are based on false beliefs, but it is clearly reasoning about *actions* (as opposed to happenings).

Cognitive psychologists have not only studied the development of the intentional stance, but also that of the *design stance*. The design stance is in this research described as the stance from which we see artifacts in terms of original intended function (design function). Some cognitive psychologists have argued that the naïve theory of rational reasoning is a precursor both to the full blown intentional stance, and to the design stance.

I am most interested in the studies on the naïve theory of rational action and the design stance. I deal with the naïve theory of rational reasoning rather than the full-blown Theory of Mind, first of all, because the naïve theory of rational reasoning is much more similar to Dennett's idea of the intentional stance, than the full blown Theory of Mind is (the Theory of Mind Stance in cognitive psychology has to be seen as a kind of higher order intentional stance). Furthermore, it is the naïve theory of rational action that helps us understand how the intentional stance develops into a theory of mind, and, more crucial to my thesis, how the design stance develops. In addition, the naïve theory of rational action is studied as applied to artifact-like things, like images on computer screens, and other "dehumanized agents".

The main purpose of this chapter is to evaluate whether Dennett's conceptualization of technical artifacts accurately describes our ordinary conceptualization of them. Do we in fact conceptualize artifacts as Dennett describes in his account of the design stance, that is to say, do we reason in terms of optimality of the design rather than intentions of a designer? Let me phrase it in a more specific question and relate it to Dennett's fourth thesis of artifact interpretation:

(1) How important is the interpretation of intention of the designer in determining the function of an artifact? Alternatively, do we ascribe functions on the basis of optimality considerations alone? (IA4)

But in addition, I will seek some global empirical answers to three other questions, that correlate to the three other theses of artifact interpretation as introduced in section 5.3 of chapter 5:

(2) How broadly do we ascribe attitudes to technical artifacts? (IAI; thesis of attitude generosity)

(3) Is the design stance related to the intentional stance, and if so: how? What stance has primacy? (IA2; thesis of design primacy)

(4) Are function ascriptions to technical artifacts similar to function ascriptions to biological items? (IA3; thesis of function generality)

In order to answer these four questions, I will first discuss the relevant findings in the psychological literature. What I will be especially looking for is how well Dennett's description of the stances corresponds with empirical reality. It will turn out that the intentional stance performs very well, whereas the design stance does not.

The intentional stance and the design stance are compared in their empirical performance for a number of reasons. First, the successful test of the intentional stance shows that it is possible and useful to test a stance on the basis of the empirical criterion. Secondly, it helps me enforce my claim that empirical research can indeed give positive strength to an account of stances like Dennett's. The research helps us to understand better how the intentional stance works (e.g. the role of rationality) and shows how much we rely on the intentional stance. Thirdly, the comparison shows that the design stance scores considerably bad, a conclusion I can only draw by comparison. Dealing with the intentional stance also allows me to show that the empirical design stance cannot be easily separated from the empirical intentional stance, pointing at an intentionalist account of the design stance rather than an optimality account.

I will close this chapter with a discussion of the value of this empirical research for philosophy. What does this developmental and cognitive research tell the philosopher? If the developmental literature goes against our philosophical theories, should we revise our theories?

6.2. Seeds of action and design: the naïve intentional stance

Although children will only be able to pass the false belief test at the age of four or five, there are some remarkable developments in the child's theory of mind earlier in its development that are worth mentioning. Two year olds are able to reason in terms of pretended items. Eighteen month old children will complete a failed action of another person, and are sensitive to speaker intentions when they learn words. Eighteen month olds also acknowledge that their own desires may be different from the desire of an adult (for references, see Griffin and Baron-Cohen: 2002, 86-88). And twelve month olds show clear indications of a simple theory of mind, such as declarative pointing, gaze following and social referencing. These abilities are taken to be clear indications and crucial parts of an early theory of mind.

At the age of twelve months, perhaps even earlier, children are able to interpret abstract, animated pictures as agents. They will ascribe goals or actions to such agents, perhaps even beliefs[59] and desires, and will be able to form expectations with respect to their future actions.

It is not clear whether this ability is a precursor to the intentional stance (theory of mind), or if it is a sign of a very early theory of mind at an early age (cf. Csibra and Gergely: 1998). For some psychologists, this naïve theory of rational action[60], is (also) a precursor to the design stance, a stance that children are able to take only much later in their development, at the age of six or seven (see below). In this section I will report the most interesting findings about this naïve theory of rational action and some further relevant developments in children's teleological reasoning.

Research on early theory of mind has focused on agency-detection and reasoning about agents. There is growing evidence that children distinguish agents from physical objects at a very early age, and that they have different expectations about the future actions of agents than about the movements of physical bodies. If this is true, children have a special "stance" for physical bodies, and one for agents.

How do children recognize agents? The literature mentions a number of perceptual agency clues. These clues are used by children to recognize agents and reason about their actions somewhere between the sixth and twelfth month of age – before they learn a language (cf. Griffin and Baron-Cohen: 2002, 89).

The most important way to recognize an agent is that an agent can move "by itself", without being physically "pushed" or "pulled" by an external force. This is called 'self-propelled motion'. Furthermore, an object can "react" to something in its environment without being directly caused to move: 'causation at a dis-

[59] Children of this age will attribute beliefs to such entities on the basis of their *own* beliefs about the world. E.g. if the child sees a barrier, it will ascribe a belief to the agent that it sees the barrier too.

[60] This simple intentional stance is called 'the intentional stance', 'the teleological stance', and the 'naive theory of rational action' in the empirical literature (cf. Csibra, Gergely et al: 1999, 241). I prefer to use 'naïve theory of rational action'. 'The intentional stance' is misleading, as it is not clear whether the primitive intentional stance can be equated with the full blown intentional stance. 'The teleological stance' is misleading, as some other authors (e.g. Kelemen: 2004) use this very same term for a (possibly) different kind of stance in which children see items as *for* something.

tance'. Agents can make jumps by themselves, they can 'decide' to move to another location, etc. (Gergely, Nadasdy et al: 1995, 167-168).

More in general, agents can move in a different way than we usually expect of physical bodies. From the external perspective, agents 'resist' the laws of nature, for instance, a bird can 'resist the law of gravity' by flying up instead of falling off a roof when pushed. Agents can jump over obstacles, can suddenly stop moving, can chase another object, can retreat. These are all kinds of movements that we do not expect of physical bodies. Dretske makes a distinction between *happenings* and *doings*. Happenings have external causes, whereas doings have internal causes (Dretske: 1988). We may say that agents are able to *do* something. Their movements have internal causes, and it is this feature that triggers in us a different way of predicting their behavior.

Let me emphasize that, to be sure, every agent obviously is a physical object. But not every physical object is able to perform the movements that agentlike-physical objects are able to perform. Rocks do not fly, and billiard balls will not avoid crashing into another object. What we determine, then, is whether we have to do with a normal physical object, or with a physical object capable of agent-like behavior.

A second clue of agency is that actions have a so-called 'equifinal structure': agents can choose different ways to accomplish their goals, and we can infer a goal from the different actions an agent performs to accomplish that goal when placed in different environmental conditions. If an agent wants to reach the other side of the room, it can just walk or roll there if there are no obstacles, but it will jump over a barrier if it gets in its way. By contrast, a physical object headed towards the other side of the room would just bump against the barrier and stop moving.

Recognition of agents, and distinguishing agents from normal physical objects, is obviously important in reasoning about the behavior of agents and objects, and in predicting their behavior. When we theorize about what an agent will do next, we will have to reason about its goals (its 'inner' states and drives), and reason about what it will be able to do given its capacities (can it jump? Can it fly? How smart is it?). Theorizing about what an agent wants to accomplish is a good way to infer what it will do next.

Now, recent cognitive psychological research strongly suggests that very young children indeed reason differently about agents than about physical objects. They will use their 'knowledge' of the assumed goals of an agent, to form expectations about its future actions, and predict its future behavior. These children will, furthermore, expect an agent to take the most rational action available to accomplish its goals. Their reasoning is constrained and guided by a rationality assumption.

What kind of research is performed to prove these claims? Usually, researchers use the so-called and well-established violation-of-expectancy paradigm for young children (these types of studies are also called 'habituation studies'). By recording the time the children look at a certain stimulus, they infer whether the behavior of the stimulus is what the child expected: short looking times indicate that the child is not surprised by what it sees, whereas longer looking times indicate that the child's expectations are violated.

The research design of these agency-studies is as follows. Researchers first present the children with one or more stimuli: one or more agents. They make sure children recognize an object as an agent by letting it perform typical 'agent-behavior' (self-propelled movement, irregular behavior etc.). This is called 'habituation': letting the child get used to the object being an agent in stead of an ordinary object. Then, the researchers present the children with these 'agent(s)' in a number of different circumstances in order to investigate its expectations about agent behavior.

Let me clarify with two example studies (see Figure 3). Csibra and Gergely habituated the children (one year olds) to the events in the first columns, and subsequently showed them two different outcomes. In experiment (a), a small yellow ball jumps over a barrier, landing close to a big red ball. The small yellow ball 'wants to be with' the big red ball. This is not behavior we would expect from a physical body – it is typical agent-behavior.

After habituation, children were confronted with two different outcomes. First, one in which the barrier is removed, but the small ball still makes the jumping-movement. Secondly, one in which the barrier is removed, and the small yellow ball rolls straight to the big red ball. The second outcome is evidently the most rational action. Indeed, children tend to look longer at the first (incompatible) outcome – which indicates that they would have expected the small yellow ball to go straight to the big red ball. (Csibra, Gergely, Biro, Koos, and Brockbank: 1999).

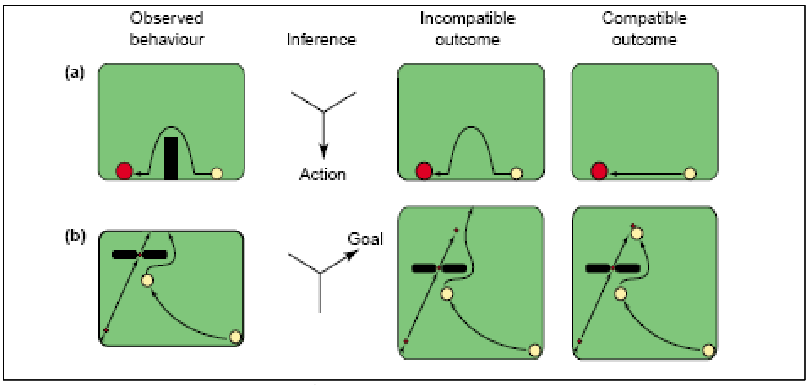


Figure 3: two violation-of-expectation studies with respect to actions (pictures are from Gergely and Csibra: 2003, 288)

Similarly, in experiment (b), the children were habituated to an event in which a yellow ball moves towards a moving black dot. When the black dot disappears through a hole in a barrier, too small for the yellow ball to enter, it apparently takes a 'detour' and seems to follow his 'chasing' movement. The end result cannot be seen. In the two outcome events, the screen is larger, so that children can see what happens next. In the first outcome event, the yellow ball, after the detour, approaches the black dot, but proceeds in a different direction before it reaches the black dot. In the second outcome event, the ball approaches the black dot until it touches it. The second outcome is the expected outcome. Indeed, children look longer at the first outcome, which indicated that they expected the yellow ball to chase the black dot until it reached it (Csibra, Bíró et al: 2003). They interpreted the habituation event as a chasing-event – how else to make sense of the behavior of the yellow ball? – and expected the yellow ball to go for the black dot.

An important feature of this 'intentional reasoning' is that the children have to assume that the agent they see is a rational agent. Without a rationality assumption, there is no reason to choose outcome (b) over (a). The rationality assumption makes the children look for and expect the most rational action (Gergely, Nadasdy et al 1995, 172)[61], given the circumstances the agent is in. The

[61] Some authors (e.g. Gergely, Nadasdy, Csibra, and Biro: 1995) argue that apparently children are able to reason about actions, and predict agent's future actions, without having a proper concept of belief (which they only acquire at the of three, or later). It is debatable whether the primitive intentional stance is properly viewed as 'the real' intentional stance: it is not at all

rationality assumption has two main functions: it is a criterion of 'well-formedness' for mentalistic action explanations, and is a inferential principle that guides and constrains the construction of such interpretations (Gergely and Csibra: 2003[62]).

Several variants of these studies have been performed. They all indicate that children infer goals, actions, beliefs, desires, and intentions from objects that have been depicted as agents. They expect such agents to perform rationally. Such studies indicate "that at least by 9 months infants can (a) attribute goals to observed actions; (b) do so even if the agents are unfamiliar abstract entities that lack human features; (c) evaluate the relative efficiency of the goal-approach in relation to the situational constraints on actions; and (d) if the relevant environmental constraints change, they expect the agent to *modify* or *change* its means action adaptively to achieve efficient goal-attainment in the new situation" (Kiraly, Jovanovic et al: 2003, 754).

As I remarked earlier, there is still a lot of debate about the exact status of this naïve intentional stance. Certainly children of a very young age show sensitivity to agent-behavior, which may be comparable with and to be distinguished from children's naïve theory of physics. Perhaps it shows that intentional reasoning is in fact possible before the central concepts ("belief", "desire", "goal", "intention") are fully developed. Arguably, it shows that children are able to attribute agency to *abstract* entities[63]. But certainly it shows that human beings have a strong tendency to reason in (naïve) intentional or teleological terms about agents (animism).

clear that children younger than 12 months old ascribe beliefs and desires to agents. But if the authors are right, this means that the intentional stance 'works' without the concept of belief, a thesis that has been challenged by some authors (e.g. Ratcliffe: 2001, Slors: 1996).

[62] This is a rather circular definition. We may get a more precise idea when we look at the (still circular) details: "the principle of rational action presupposes that (1) *actions* function to bring about future *goal states*, and (2) goal states are realized by the most rational action available to the actor within the *constraints of the situation*. Thus, the principle asserts that a mentalistic action explanation is well-formed (and therefore acceptable) if and only if, the action (represented by the agent's *intention*) realizes the goal state (represented by the agent's *desire*) in a rational manner within the situational constraints (represented by the agent's *beliefs*)" (Gergely and Csibra: 2003, 289[62]).

[63] Some explain the sensitivity to religion on the basis of these habituation studies, see, e.g., Kelemen: 2004.

Some have hypothesized (e.g. Kelemen: 1999b; Kelemen: 2004) that the naïve theory of rational action is a precursor to a different kind of stance, the teleological stance, or, later, design stance, in which objects are seen as being made 'for' something. The teleological stance, the design stance, and their relation to the naïve intentional stance are the subjects of the next sections.

6.3. The development of the design stance

Recently, cognitive psychologists have also taken an interest in the development of our functional reasoning about artifacts or what they call, after Dennett, the design stance. By and large, this development seems to be as follows:[64]

(1) At the age of three, children distinguish between self-serving and other-serving functions (teleological stance).

(2) At about the same age, children distinguish between natural (i.e. physical) kinds, biological kinds, and artifact kinds.

(3) At the age of three, four or five, children have a theory of mind. They are able to reason about actions in terms of attitudes, and will pass the false-belief test (full blown intentional stance).

(4) At the age of four or five, perhaps six, children frame the function of an artifact in terms of original intended function (design stance).

(5) At the age of six or seven, children are able to integrate the design stance in their practical reasoning about artifacts. First signs of 'functional fixedness' (practical design stance).

(6) At the age of nine or ten, children will reserve the notion of function to artifacts. They will stop ascribing functions to natural non-biological kinds, and animals (exclusive design stance).

Why are cognitive psychologists interested in the design stance? First, there is the well documented phenomenon of *functional fixedness*. Adult human beings tend to become "fixed" on the design function of an artifact, and are, as a result, very bad in using an artifact for an atypical function (Defeyter and German: 2003). This tendency to "see" artifacts strictly in terms of intended function, has

[64] See, e.g. the studies of Kelemen, Keil, German, Johnson, and Defeyter, to which I will refer more fully in the remainder of this chapter.

lead researchers to hypothesize that the core of the concept of an artifact is, for adult human beings, the original intended function (cf. German and Barrett: 2005). These researchers have subsequently hypothesized that human beings are endowed with a 'design stance' that is responsible for our framing of artifacts in terms of intended function.

Second, and related, the design stance plays an important role in so called essentialistic psychology, an important branch in cognitive psychology. Essentialistic psychology says that people categorize entities according to their essence. There is some converging evidence that people indeed categorize *natural kinds* and *animal kinds* according to their essence. Natural kinds (which means in these contexts: physical kinds), then, are conceptualized in terms of their causal properties. Animal kinds are conceptualized in terms of their origin (see, e.g., Keil: 1989).

Technical artifacts pose a problem for the essentialistic psychologist. Until recently, it was thought implausible that artifact concepts have 'essences', or "cores', like natural kinds and animal kinds. An artifact kind is usually considered an unnatural kind. For instance, unlike natural kinds, artifact kinds do not have sciences around them, require reference to human intention, and do not have unique paths of origin (cf. Bloom: 1996, Keil: 1995, 235, also for a criticism of this view). (Note that these psychologists, unlike Dennett, treat biological kinds and artifact kinds as different kinds.)

However, the results of the functional fixedness studies have inspired essentialistic psychologists to find proof that artifacts are, like natural and animal kinds, categorized on the basis of their essence. The essence of an artifact, then, is assumed to be its original intended function, which is supposedly grasped by means of the design stance. Research has indeed indicated that people seem to categorize artifacts on the basis of intended function. More precisely, when having to choose between naming the artifact after the intended function, and other properties (such as accidental function, or physical appearance), people prefer to name the artifact after the intended function, as long as that function is feasible (Matan and Carey: 2001, see Malt and Johnson: 1992 for an anti-essentialistic account (see 6.5.1)).

A third reason for the interest in the design stance is that it seems to require a quite complex form of higher-order intentional reasoning, that is, forms of reasoning that invoke second-, third- or even higher order references to mental states. The design stance, then, would need a reference both to the intention of the designer, and to user intention (“the designer intends me to intend to use the artifact so-and-so”). Thus, higher-order intentional reasoning is only possible *after* children *master* the intentional stance (i.e. have a theory of mind). This thesis is not shared amongst all researchers of the design stance. I will discuss the higher-order reasoning thesis in section § 6.5.2.

6.4. Three models of the development of teleological reasoning

There are several hypotheses about the development of teleological reasoning. Some researchers think that human teleological reasoning is gradually refined during child development. They think of the development of teleological reasoning as a fork-like structure, in which the naïve theory of rational action slowly evolves into the two types of teleological reasoning we are familiar with: the intentional stance for agents (theory of mind), and the design stance for artifacts (framing of artifact function in terms of original intended design).

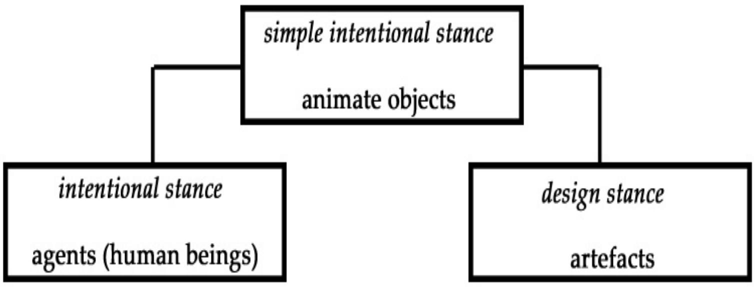


Figure 4: the refinement model of teleological reasoning (Kelemen)

We could say that they learn to distinguish goals (or actions) from functions. At the same time, they learn to restrict their application of the stances to the appropriate categories: actions belong to agents, and functions belong to artifacts. I will call this “the refinement model of teleological reasoning” (Figure 4). Deborah Kelemen defends a model like this. I will discuss it below (6.5.3)

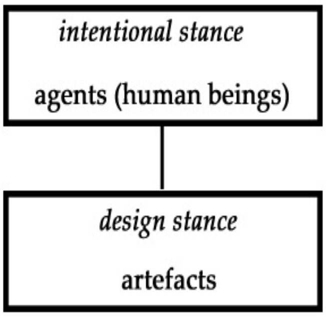


Figure 5: The linear model of teleological reasoning (German and Johnson)

A variant (Figure 5) on the model is what I call the linear model of teleological reasoning. According to this model, the design stance develops out of the intentional stance. This model is most clearly defended by German and Johnson (2002, see also section 6.5.2).

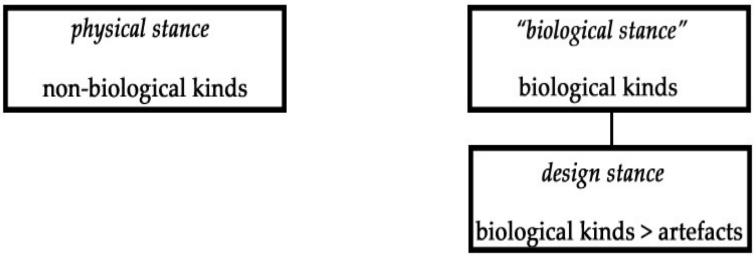


Figure 6: the autonomous model of teleological reasoning (Keil)

The third model worth mentioning is one in which the design stance is not framed in terms of the intentional stance at all (cf. Keil: 1989). According to this model, which I call the autonomous model of teleological reasoning, the design stance derives from a – probably innate – structure to distinguish biological kinds from natural kinds (a "biological stance" as contrasted with a physical stance). From this paradigm, the design stance is meant as a structure for biological kinds, that, in our culture, is also used for technical artifacts (Figure 6).

Further research has yet to show which model (if any) is correct, but Kelemen's recent evidence strongly supports her own refinement model. Keil's older findings can be plausibly reinterpreted as to fit her paradigm. And the German/Johnson model is not really competing with Kelemen's view – as they operationalize the design stance slightly different than Kelemen (2004, see also footnote #71). I will therefore largely rely on Kelemen's findings. Let me first try to give a rough idea of the development of the design stance, on the basis of – sometimes conflicting– evidence[65].

6.5. X is for Y: The development of the design stance

6.5.1. Categorizing artifacts in terms of original intended function[66]

The design stance is a stance that develops over the period of several years. It is not yet clear when exactly the first signs of a design stance are present, and it is also not yet clear when children possess a full-blown design stance. Some researchers claim that the design stance develops in a rather short period, when children are five or six years of age. Others, such as Deborah Kelemen, have claimed that it starts much earlier, around the time that children learn to reason in terms of purposes (when they have a naïve theory of rational action, that is, before the age of twelve months), and is only fully present at 9 or 10 years, when they restrict their application of the design stance to artifacts and biological traits. I will mainly rely on Kelemen's studies, because she has tried to give an integral overview of the development of the design stance for its own sake.

Kelemen's studies indicate that children of two to three years already show a sensitivity to intended function: "as early as age 2½, children need only one exposure to an adult intentionally using a novel tool to rapidly and enduringly construe the artifact as 'for' that purpose rather than any arbitrary activity it

[65] The conflicting evidence is largely due to the fact that researchers use different criteria and operationalize the design stance differently. This, in turn, is due to the fact that they often have different research goals and programs. For instance, German and Johnson are most interested in the practical reasoning attached to the design stance (functional fixedness), whereas Kelemen focuses on more 'theoretical' teleological ascriptions.

[66] The following paragraphs draw heavily on *Optimality vs. Intent: Limitations of Dennett's Artifact Hermeneuticss* (under review), written by Krist Vaesen and myself.

physically affords" (Casler and Keleman: 2005, 478). Casler and Kelemen believe that the sensitivity to intended function is an important step in the cognitive development of human beings. Contrary to captive monkeys, human children learn from, and imitate, the way adults use an artifact, and adopt the intentional use of an artifact as its 'right' use[67].

At the age of five or six[68], perhaps earlier, children will start to classify artifacts on the basis of original intended function (German and Johnson: 2002, Kelemen: 1999b, Matan and Carey: 2001, Defeyter and German: 2003). Matan and Carey found that an artifact that was presented as designed as a teapot, and was now being used to water plants, would be judged a tea pot by adults and six year olds, but not by four year olds. In another experiment, Kelemen showed five year old children an artifact that was designed to dry cloths. The subjects were confronted with situations in which the artifact was used in a different way, namely, as a back stretcher. Children would judge that the artifact was not for back stretching, whether this new use was a one-time accident, a one time intentional use, or repeatedly intentional use.

A study by Gelman and Bloom (Gelman and Bloom: 2000) points out that children even from the age of three to five conceptualize something as an artifact when they believe it was intentionally designed. For instance, a piece of paper with the form of a hat was more frequently interpreted as a 'hat' when subjects were told that it was intentionally designed as a hat, than when the hat-form was caused by a non-intentional process (e.g. a car ran over a newspaper). In both conditions, the piece of paper fulfills the function of hat in the very same optimal way. So, if optimality would be guiding in subjects conceptualization of artifacts,

[67] Kelemen reports research that suggests that children of two to three years classify artifacts mainly on the basis of physical appearance, like shape (reported in Kelemen 2004). She suggests, however, that children may infer intent from shape.

[68] Kelemen (Kelemen: 1999b) reports categorization on original intended function already at the age of four/five. German and Johnson suggest that Kelemen's studies are biased towards original intended function, because the new use conditions were framed in accidental language. In their experiments, they found that if original use is intentionally changed, children were not more likely to pick original function over current use (286). Kelemen has, in response, suggested that there may be a difference in artifact categorization, and stating what some artifact is 'really' for.

we would expect the number of subjects that consider the paper object a hat to be equally distributed under the two conditions[69].

Another study certainly worth mentioning was done by Deborah Kemler Nelson and colleagues (Kemler Nelson, Herron et al: 2002, cf. Kemler Nelson: 2004). They researched the categorization of broken objects, in order to find out whether children classify artifacts on the basis of intended function rather than current (optimal!) function. They investigated two kinds of dysfunctional objects: accidentally dysfunctional items (e.g. a fork with broken tines) and intentionally dysfunctional items (e.g. a scissor where the finger holes were glued permanently together). Accidentally dysfunctional items were classified as artifacts significantly more often than intentionally dysfunctional items. The authors conclude that inferences about functions intended by object designers guide the way artifacts are categorized. If they are correct, they seriously undermine Dennett's optimality account of functions. According to Dennett's paradigm in which we have to assume that the object is not malfunctioning, a broken item would not even be *viable* for design interpretation.

At the age of six or seven, children start to show the signs of functional fixedness, suggesting that original intended function is at that time at the heart of their perception of and reasoning about artifacts. They not only "theoretically" frame an artifact in terms of original intended design, but will actively and practically use this knowledge (or bias, if you wish) when asked to solve specific problems with the artifact. For instance, in one study Defeyter and German showed that six-seven year old children were slower than five year olds in solving a problem by using an artifact in an atypical way. (Defeyter and German: 2003, see also German and Johnson: 2002). We may perhaps say that the functional fixedness test is for the design stance what the false belief test is for the intentional stance.

Paul Bloom is perhaps the most explicit defender of the thesis that intended original function is indeed at the core of our concept of artifacts.

> We infer that a novel entity has been successfully created with the intention to be a member of artifact kind *X* – and this is a member of artifact kind *X* – if its appearance and potential use are best explained as resulting

[69] Thanks to Krist Vaesen to drawing my attention to this study.

> from the intention to create a member of artifact kind *X*" (Bloom: 1996, 12).

Several other studies further indicate that the notion of (artifact) function is conceptualized in terms of original intended function (e.g. Jaswal: 2006).

But the empirical debate is certainly not unequivocal. Malt and Johnson question whether intended function is really an essential property in categorization. They claim that it is neither a sufficient, nor a necessary condition. For instance, artifacts that are dubbed to be a boat, but that do not fulfill physical expectations we would have of a boat, were not called boats. More studies indicate that physical features are sometimes given more weight than intended function. Chaigneau (2002), for instance, also argues that original intent must be supported by the artifacts' structure. For instance, in one scenario the experimenter intended a certain object to be a mop and used it as a mop, but the object was a bundle of plastic bags attached to a four foot long stick. Most subjects did not regard this object as a mop, despite that it was its intended function and that it was used as such (similar results are had from, e.g. Hampton: 1995, Landau, Smith et al: 1998, Baldwin: 1992, Sloman and Malt: 2003).

It is not so clear what to make of these results. Shape is obviously an important cue to interpreting *optimal* function, rather than intended function as it tells us something about what the artifact is plausibly able to do well. But the results can be interpreted in different ways. Diesendruck, Markson & Bloom (Diesendruck, Markson et al: 2003) have argued, for instance, that the shape bias is in fact reducible to an intentional bias. Shape, they argue, is an excellent cue to intent (rather than optimal function). One way of testing this is to develop an experimental set-up which parallels our common sense observation that containers are categorized differently from the objects they contain, despite the fact they have the same shape. The authors conclude:

> Given that in most real-life cases the creator's intent is not readily available, children must rely on cues to the intent. An object's shape and function are examples of these cues. An object that has the specific shape of a chair and that serves primarily for one person to sit on was likely created to be a chair. (Diesendruck, Markson, and Bloom: 2003, 168)

So, Diesendruck and colleagues think that the intent of the designer is a `more conceptually central' property (164) than mere function (in the sense of performance) or shape (similar conclusions are found in Gelman and Bloom: 2000).

I should mention that the empirical studies under discussion aim to prove (or disprove) the importance of reasoning about designer intent in the categorization of artifacts, and are not designed to prove or disprove the optimality account per se. So the research is not conclusive with respect to the question whether the optimality account is empirically wrong. Yet, the studies do show that the empirical design stance is highly intentionalistic and this does make the optimality account highly implausible. *If* optimality considerations play a role, we may at least expect that intentional reasoning has an important position in it – much more than Dennett's optimality account admits.

The results, then, are not unequivocal, and probably subjects weigh many factors when classifying such complex objects such as artifacts, including context, design history, physical features, and use. Still, it is rather plausible that the design stance normally develops towards a conceptualization of function in terms of original intended function and that that factor weighs significantly heavier in categorization than others– as long as the artifacts' form and structure do not make it impossible to fulfill that function. Indeed, Malt and Johnson admit that original intended function has a heavy weight, and that people will categorize artifacts on the basis of their original intended function if that function is feasible (Malt and Johnson: 1992).

6.5.2. Does the design stance require higher order reasoning?

I think we may conclude from the empirical research discussed that the design stance involves a recognition and interpretation of the intentions of the designer who produced the artifact, which means: taking the intentional stance. Taking the design stance, then, requires some pretty advanced higher-order reasoning about the intentions of the designer, something which, according to German and Johnson, is applied with respect to artifacts only from the age of six, or even seven (years after the capacity to theorize about the mind).

German and Johnson give two explanations for the late development of the full blown design stance. First, it may be hard for five year olds to reason about origins. Secondly, the full blown design stance may require higher-order intentional reasoning.

> we suggest that understanding design is more complex, in representational terms, than understanding a simple goal-directed object use. This proposal stems from the idea that the notion of "intentionally made for purpose x" involves coordinating two mental states: first, that of the maker, and second, that of a subsequent user. One way of capturing the notion of design, therefore, is as a recursive mental state, as in "the maker intends that 'the user intends that x'" (German and Johnson: 2002, 297)

Children are known to develop the ability to reason about recursive mental states relatively fairly late (297). This would explain, according to German and Johnson, why children are only able to take the design stance at such a late stage[70].

6.5.3. Objects of teleological explanations: is there a promiscuous teleology?

We have seen that children categorize artifacts on the basis of original intended function approximately from the age of five, or six. But at that point, children do not restrict their application of the design stance to artifacts. Until the age of nine, they ascribe functions and purposes to all kinds of entities: biological traits (wings), whole animals (lions), and non-living natural kinds (mountains). For instance, children claim that mountain peaks are for climbing, that pointy rocks are for scratching, that sand is grainy so that it will not blow away (self-serving), or that sand is grainy so that animals can easily bury their eggs in it (other-serving) (Kelemen: 1999c). Other-serving teleological explanations come a little bit later in development. Preschoolers will assent to the thesis that pencil shavings are made for something. It is only at the age of nine, at school, that they learn that the design stance is not properly applied to non-living natural kinds and animals (Kelemen: 1999a; Kelemen: 1999b). Until then, children also endorse the idea that clouds that stop raining should be repaired or replaced – or that it may be "time to get a new mountain" (DiYanni and Keleman: 2005).

According to Kelemen, both the design stance and the intentional stance derive from the more primitive teleological stance, the stance from which

[70] Kelemen (2004), however, questions whether the recursive reasoning thesis is plausible: the second order reasoning can be reconstrued as a simple first order reasoning (maker intends that x does y), and complex second order reasoning is present in children from age three or four. But Kelemen agrees that the interpretation of intentions is crucial to the design stance, and to the perception of artifacts.

activities are seen as done for a purpose. When children encounter artifacts, they often do so in the context of someone using the artifact for a purpose – the ascription of purposes to an artifact, then, derives directly from the inference of the purposes of users, or, later, designers. These findings has brought Kelemen to hypothesize that children are "promiscuous teleologists": they generously give teleological explanations to a wide range of entities[71]:

> "[PT] argues that purpose-based explanations are generally compelling to people because teleological reasoning is derived from a mode of thought that, due to our evolution as complex social animals, comes easily to us – intentional reasoning. Specifically, it seems possible that the tendency to explain objects in terms of a functional purpose develops from our bias to explain the behavior of agents by attributing "mental" purpose... (Kelemen: 1999c, 27)

A teleological explanation, then, is an explanation that assumes that objects or events occur for a purpose (Kelemen: 1999c). Children are "intuitive theists", "predisposed to construe natural objects as though they are nonhuman artifacts, the products of non-human design" (Kelemen: 2004, 295)

In her promiscuous teleology paradigm, Kelemen explicitly links the primitive design stance to the naïve theory of rational action as investigated by Csibra and Gergelny:

[71] Let me note that Kelemen's results differ from the established paradigm, "selective teleology", that says that children limit their teleological explanations to biological traits and artifacts (Keil, Atran, cf. 6.4). I have called this paradigm the autonomous model of teleological reasoning. The selective teleology paradigm relies on empirical findings that children can distinguish between biological traits and artifacts, and that they distinguish artifacts from biological traits on the basis of the other-serving nature of artifacts, and the self-serving nature of biological traits. That is, artifacts fulfill functions for external agents, whereas biological traits fulfill functions for the organism to which the trait belongs (cf. Kelemen: 1999c). The selective teleology paradigm suggests that children have an innate, specialized module for recognizing biological kinds from other natural kinds. At a later age, and influenced by a culture full of technology, children apply this innate teleological module to artifacts as well. The design stance, according to this view, derives 'accidentally' from a bio-design stance, and develops relatively autonomous from the intentional stance. Kelemen has argued that her paradigm accommodates the Keil/Atran findings, while explaining more findings (amongst which the teleological stance, and her experimental data) that apparently contradict the selective teleology paradigm (Kelemen: 1999a).

> "Between 6 and 9 months, babies construe animate objects as goal-directed agents (...) and by 12 months, infants use this mode of construal to predict a novel object's future behavior (...) This rudimentary teleological stance is then rapidly embellished as children notice that agent's goal-directed activities are often focused upon objects that are employed as means to an end..." (Kelemen: 1999b, 245)

Kelemen's paradigm suggests that for children the difference between purposes-as-goals (intentions) and purposes-as-functions (functions) is much less clear than it is for adults. Adult human beings make a strict division between goals and functions, goals being the purposes of agents, and functions being the purposes of artifacts (or biological traits), derived from the goals of agents. Children attribute functions to a wide range of entities, including animals and non-living natural kinds, and they attribute goals to abstract agents, including artifacts. It would be interesting to see whether young children also attribute functions to human agents (mommy is made for playing with), and how their intentionalizing of artifacts develops (do children stop thinking in intentional terms when they develop the design stance?).

Kelemen's work strongly suggests that human beings are very much focused on intentions, true intentionalists, and that it takes school to have them learn to restrict their intentionalist explanations of phenomena to the "appropriate" categories. As such, Kelemen's work would speak in favor of a design stance that is as liberal as the intentional stance, a stance that seeks purposes in a highly intentionalist way.

6.6. Functional reasoning is intentional reasoning

I started this chapter with one main question and three other questions:

(1) How important is the interpretation of intention in determining the function of an artifact? Alternatively, do we ascribe functions on the basis of optimality considerations alone? (IA4)

(2) How broadly do we ascribe attitudes and to what extent to artifacts? (IA1)

(3) Is the design stance related to the intentional stance, and if so: how? What stance comes first? (IA2)

(4) Are function ascriptions to technical artifacts similar to function ascriptions to natural items? (IA3)

We may, I think, safely conclude that the cognitive psychological research supports a large part of Dennett's theory of *action* interpretation (question 2) – if we compare Dennett's (generous) intentional stance with the naïve theory of rational reasoning. Very young children ascribe simple attitudes to computer images on screens, abstract items that have little resemblance to human beings. Even stronger, it seems that children start their lives with an abstract sense of agency. The naïve theory of rational reasoning is a very generous stance, that probably underlies many other intentionalistic (and otherwise teleological) ascriptions.

Action interpretation, we may conclude, works more or less along the lines as Dennett has described it. The research gives empirical support and more detail to the rationality assumption, and gives more insight in the inferences made in interpretative reasoning (as holistic reasoning). The (generous) intentional stance can, furthermore, be very well used for non-human beings, and is used as such. In addition, the research gives concrete details to the thesis that the intentional stance gives us extra predictive value, and explains this: agents are harder to predict than physical objects, and the intentional stance is, conveniently, a very good strategy to predict and explain agent behavior.

The cognitive psychological research on the *design* stance, however, is not so favorable for Dennett. Original intended design plays a very important role in our conceptualization of artifacts, even to the extent that we get "fixed" on it, and find it hard to use an artifact differently from its original intended design. There is, it seems, no design stance without a prior intentional stance.

The third question, how the design stance and the intentional stance are empirically related, is still under heavy debate, but some early and careful conclusions seem to be supported. First, the naïve theory of rational reasoning plays an important role in the development both of the full blown intentional stance, and the design stance. Children's high sensitivity to agency makes them see items as created for a purpose by a more or less intelligent agent. Goals and functions are virtually indistinguishable for the young child. How exactly the stances are related is uncertain, but what is clear is that the design stance is one of the last teleological stances to arise in development. *If* there is a relation of priority between the design stance and the (generous and strict) intentional stance, then the design stance comes last.

The answer to question 4, whether function ascriptions to artifacts are similar to function ascriptions to natural items, depends on whether we take the teleological stance (*X* is for *Y*) as our reference, or the design stance. The teleological stance is applied generically to technical items, biological items and even such things as mountains until the age of nine or ten, when children learn about the theory of evolution at school. We may therefore carefully conclude that, were it not for formal education, human beings might approach technical artifacts in just the same way as biological items. Our “natural intuitions”, then, tend towards a generic account. Not generically optimal , but generically intentional!

If we look at the design stance itself, a different picture arises. Function ascriptions to artifacts are made in terms of original intended design, where original intent is clearly seen as *human* intent, not ‘natural intent’. Still, the design stance in this form only arises late in development, and it has, as far as I know, never been tested whether and how children take the design stance towards natural items. In cognitive psychology, the design stance is *defined* as an artifact-only stance, so it is hard to draw any conclusions about the genericness of functional ascriptions on the basis of this research on the design stance.

It seems that our cognitive apparatus until rather late in development, takes function ascriptions to biological items and technological artifacts indeed to be the same. What is safe to conclude, I think, is that human beings have a natural tendency to see and explain things – both biological and technological - in teleological terms. Whether or not the design stance is a specific stance targeted at artifacts, not to biological items, and how it is related to the teleological stance, is less certain and rather unclear.

Concluding, when it comes to the interpretation of artifacts, their categorization and problem solving with artifacts, human beings focus on, or are biased towards, intended, original design. Human beings are intentionalists, and their intentionalism lies at the basis of the further development of their intentional reasoning, as well as their reasoning about functions. Sensitivity to intended design arises quite early in development, from the age of four, only to have really settled in the minds of children at the age of six, or seven. Knowledge of the intended original function may be overruled by dysfunctional shapes or structures, but only in extreme cases. The "intentional fallacy", we may say, is just a fact of human life.

Given the pervasiveness of intentional reasoning, also when it concerns functional items, we may speculate that there are good reasons to have intended original design at the heart of artifact concepts. Reasoning from Dennett's standards: given the pervasiveness of intended design in our concept of artifacts, we must assume that there is a good reason (be it free floating) to do so.

So what could be reasons why human beings develop this design stance? Is the design stance a bias in human being's reasoning about artifacts (as the term 'functional fixedness' suggests)? If so, is it a useful bias? As far as I know, there are no straightforward explanations. Casler and Kelemen (2005) indicate some of the positive aspects of the design stance:

> The phenomenon of functional fixedness is widely recognized and has downsides as well as advantages. "[A] downside to representing artifacts in terms of intended design is that it may inhibit us from violating that function when an alternative use might be advantageous ('functional fixedness') (...) This is a small price to pay, however, for the powerful benefits of the design stance. It significantly underpins our capacities as the most organized and efficient species of tool-using problem-solvers. Even more distinctively, it permits us to manufacture and use an astonishing diversity of specialized tools, since our stable representations of what existing objects are 'for' allows us to innovate and use new objects for new tasks" (Casler and Keleman: 2005, 472)

Indeed, it is not hard to see that the design stance may help us, as a useful shortcut, for the use of artifacts. That some artifact was deliberately created by a designer, gives us reason to think that we can use it for that intended purpose. What the artifact was created for is what it is most likely to be good at for the simple reason that *someone else* already thought about the best way to use it. Sticking to the intended function prevents us from inventing the wheel over and over again, or using the artifact for suboptimal functions. But Casler's and Kelemen's claim that the design stance is a condition for innovation and creativity strikes me as rather odd, if not contradictory: we just said that alternative tool use is *discouraged* by the design stance, rather than promoted. Perhaps what Casler and Kelemen mean, is that as functions get fixed by specific artifacts, we are encouraged to think of the creation of new artifacts for new functions ("this artifact is already occupied – find your own artifact"). A washing machine is for automatic washing of clothes, so when we are looking for a way to fulfill a different function (say, washing babies), we are prone to develop a new artifact that can best fulfill this new function. The design stance, according to such a hypothesis, would not only promote the development of new artifacts, but also lead to a differentiation of artifacts, all for their own function. Unsure about whether Casler and Kelemen indeed support such an hypothesis, I find this reasoning rather far-fetched.

I take it that the explanation of the existence of the design stance is not at hand yet, and that research on the development of the design stance is meant to shed more light on the explanation of its existence. It may be an innate strategy that evolves autonomously – in that case we may have reason to think that it has, indeed, a proper natural function. Or it may have developed out of a biological stance, that happens to be applied to artifacts. In that case, the design stance rather has an accidental function as we may say. Alternatively, it may be a culture-specific stance, which is plausible given its late development. In that case, we may want to look for a more cultural explanation of the design stance.

If it turned out to be true that there is, indeed, a good explanation why people reason about artifacts in terms of original intended design, Dennett should take that reason at heart: if an intended design account turns out to be more productive than an alternative account, we may want to switch to the intended design account (he uses the same reasoning, after all, for the intentional stance!) Our strong tendency to view things in intentional terms may count as a reason to *distrust* our intuitions about design, *or* it may count as a reason to use it as an efficient methodology.

The first conclusion, then, is that Dennett's account of the design stance as it stands is *not* an empirical account. We *do* reason in terms of intentions when we reason about artifacts. Not only that, reasoning in terms of intentions, goals, and purposes is crucial to the human mind, more so than reasoning about artifacts.

Furthermore, if Dennett chooses to ignore the empirical counterpart of the design stance, he should also ignore the empirical counterpart of the (generous) intentional stance ("gelijke monniken, gelijke kappen", as the Dutch saying goes) – and that is truly a missed opportunity. As a naturalist, I would say, Dennett would better start from scientific data about design and intention, and build his theory on the basis of that data.

If Dennett would insist that his account of the design stance is a methodological, normative account, he should at least justify why he departs from the empirical account, especially if it is shown that the intentional account is predictively or practically *stronger* than, say, the optimality account.

6.7. Discussion: The relevance of empirical (developmental) research

Philosophers are often skeptical about the use of empirical material for their theories. Most philosophical theories are normative, in a broad sense. They specify some kind of ideal, be it epistemological, ethical, or methodological. And, as Hume taught us, you cannot derive an *ought* from an *is*. But an ideal is always an ideal of *something*, usually a 'correction' of it, and every normative theory has to start somewhere: the way we actually do it. It would be strange to formulate a normative theory that has no link whatsoever to empirical reality.

There is some debate about the extent in which normative theories can depart from our everyday way of doing things. For instance, in epistemology there is a hot debate going on about whether it makes sense to formulate epistemological theories that are based on ideal forms of rationality that are probably non-existent in our world (see, e.g. Stich: 1990, Stein: 1996). *Ought* implies *can*, also in epistemology. But what does "can" mean? Is it a 'can-in-ideal-circumstances', or is it a 'reasonable can'? I *could* calculate the square root of 4,3 without a calculator if I really had to, but it would take me a long time, and I would certainly be skeptical about why I am expected to do it in the first place. A similar debate is going on in meta-ethics; can we expect people to do the 'right thing' without them being motivated to do it? If the normative theory departs too much from our everyday intuitions, it may be a nice normative theory, but one that will be hard to put to practice. It would perhaps not be a normative theory at all.

So our normative philosophical theories depend at least to a certain extent on descriptive theories. Some normative theories stick relatively close to the descriptive theories (e.g. ordinary language philosophy), others make some revolutionary changes (e.g. analytic metaphysicians telling us that chairs do not exist).

In analytic philosophy, constructing theories about our concepts of things (actions, intentions, good, bad, knowledge, reason, function) has become a big business. These theories usually start from our ordinary use of terms, but they have a normative flavor because they look for inconsistencies in the proposed concept, and try redefine the concept in order to correct these inconsistencies. A concept, in that sense, is always a kind of idealization A drawback of this methodology is that the primary sources usually consist of data that are not gathered and studied systematically. It is here that I think that empirical studies on concepts and philosophical analysis could enforce each other.

The cognitive empirical research that I have discussed in this chapter, is not so different from this philosophical practice. It also tries to map how we ordinarily use certain terms, and how we classify things. But its subjects form a representative group. In fact, many of these empirical studies *experimentally* test the intuitions of their subjects with respect to certain concepts. So, philosophers and empirical researchers are, in this case, after the same thing, using a different method. So, if we want to construct a theory about a certain concept, and if we think that it should be based on the way we actually do it, this empirical data may form a wealthy and systematic source[72].

Developmental research may even challenge certain philosophical conceptualizations. For example, some philosophers believe that having a belief requires having the concept of a belief, which, in its turn, requires having a language. The developmental research that I discussed seems to point out that this is probably not the case. Children seem to reason about intentions long before they speak a language and without, arguably, having the concept of belief. And even if this were disproved by empirical research: the research *has* something to say about these issues, and is, therefore, relevant.

The results of this chapter are, I submit, not only relevant for naturalists like Dennett, but also for conceptualists. *If* we want to find out what we mean by certain concepts, we should be interested in the kind of empirical research under discussion here as a systematic endeavor to find out how we use these concepts. We could then critically compare the empirical results with the philosophical results.

[72] See, e.g. Bloom: 1996, Knobe: 2003 for a good example of how analytic philosophy and empirical science can go enforce each other; similar studies are done in the domain of ethics, e.g. see the Marc Hauser website http://moral.wjh.harvard.edu/index2.html which aims to empirically investigate moral intuitions

7 Taking a Stance as a Method

If Dennett's principle of optimality is indeed not an adequate description of our way of ascribing functions to artifacts, it becomes even more important that Dennett convincingly shows that it provides at least a good method for science. I test the optimality account in a critical case – archaeology – where it should be demonstrably stronger than its intentionalistic rivals and discuss some of the alleged advantages of the optimality account, such as predictive force.

7.1. Evaluating the stances as (scientific) methodology

My main research question is whether Dennett's theory of the stances provides an adequate framework of ascribing functions to artifacts. This can be read in two ways: empirically (do we actually ascribe functions to artifacts as Dennett describes) and methodologically[73] (is it wise to ascribe functions to artifacts as Dennett describes).

I have tried to show in the previous chapter that Dennett's optimality account is insufficient as an empirical theory as it fails to account for the importance of reasoning about intentions in functional ascriptions. This is unfortunate: the theory of the stances is significantly stronger when it is firmly grounded in empirical data – if only because as such it could almost directly be set out at least as an *everyday* method of conceptualizing artifacts[74]. In this chapter, I aim to give

[73] There is a distinction between method and methodology, method being a certain kind of procedure or technique for doing something, and methodology referring rather to the whole set of procedures and their justification. The concrete application of a stance in a specific field should then be seen as the application of a method, whereas the theory of the stances is a methodology. The distinction is not always sharp.

[74] A methodological, scientific account of attitudes does, according to Dennett, not necessarily start from ordinary use of terms; it may even result in a conceptualization of attitudes that is significantly different than in our everyday use of the language. In that respect, Dennett seems to have a certain freedom to depart from ordinary language use, which could explain and justify the weak results on the empirical score. But as I have argued before (3.8), a strong empirical foundation and a close fit between the application of ordinary attitude terms and scientific attitude terms does make the theory of the stances stronger, and lack of it should be

an answer to the methodological question: is it wise to ascribe functions to artifacts as Dennett proposes? If Dennett could make plausible that the optimality account of the design stance could indeed be a clear and helpful method to ascribe functions to artifacts, this may compensate for the weak results on the empirical score. The crucial question, then, is whether the account describes a good method for making functional claims in science[75]. Parallel to chapter 6, we can phrase the following central question for this chapter:

(1) How important should the interpretation of intention of the designer be in determining the function of an artifact? Alternatively, should we ascribe functions on the basis of optimality considerations alone?

The methodological question is relevant because Dennett himself repeatedly emphasizes that his theory of the stances should (also/primarily) be read as a methodological theory. Dennett believes that a methodological, scientific account of the intentional stance gives an interesting sense to attitude terms (Dennett: 1987e) and proposes to justify our using intentionalistic terms on the basis of a pragmatic and methodological argument ('they provide good/useful explanations') rather than on the basis of an ontological one ('they refer to real entities in the brain'). Dennett takes a different and more modest path than, for example, Fodor, who wants to *vindicate* ordinary attitude talk through science. Dennett does not think that science will vindicate the everyday sense of attitude terms, but he does think that a revised, idealized, account of attitudes can be put to good use in science. Dennett thinks that there is great strength in the interpretative stances, but he is quick to add that we cannot simply take them for granted in science. We may apply them in science, as a method, but critically and carefully. We may use attitude terminology as abstract terminology within an explanatory theoretical framework, if we make sure to distinguish them from their everyday counterparts and empirically be critical about the ontological

compensated. Preferably there is a close fit, and caution with respect to the ontological conclusions we draw from it (3.7).

[75] By scientific method I mean a method by means of which we can isolate, describe and explain certain events and processes in the natural world. The scientific methods of the intentional stance and the design stance, then, describe those processes that we wish to read as teleological. There are good reasons to evaluate Dennett's account as a method.

consequences we may be tempted to draw from it. The intentional stance is powerful, but we will have to accept its limitations.

Dennett's construction of the (generous) intentional stance as a method is an interesting move, and, as we shall see, one that seems to works rather well. In 7.2 I illustrate that by means of cognitive ethology, a scientific branch where the intentional stance (or something very similar) is used frequently to explain intelligent behavior of animals in terms of their cognitive states (what they 'know'). Animal minds are and will remind a controversial topic in science, but the explanatory force of concepts such as 'beliefs' and 'desires' is too powerful to ignore[76]. I also use cognitive ethological studies to make plausible that Dennett's thesis of generosity (that we may generously ascribe attitudes to all sorts of animals) is just actual practice in this scientific field. Together with the results on the empirical score, I believe this adds up to an elegant account of belief-ascriptions. Lastly, I use cognitive ethology to show the contrast between Dennett's pragmatic/naturalistic approach and more conceptualist approaches. Here again, many cognitive ethologists turn out to share Dennett's views.

By contrast, the methodological credentials of Dennett's optimality account of the design stance seem very low. It isn't easy to actually show that. For even if I could show that the optimality account fails to provide a good method in a certain scientific domain, I would not have shown that it isn't a good method in *every* domain we can possibly imagine. For my argumentation, I have chosen the tactic of the critical case. If I can pick a scientific domain in which Dennett's optimality account should show its worth, and if I can additionally show that it does not, I would have a strong enough case against Dennett. I have chosen the domain of (ancient) archaeology. Archaeology is a field that Dennett himself uses a lot as an example of the design stance in scientific practice, a domain in which not only the optimality account *should* be strong, but a domain in which its main competitor, intentionalism, should be especially weak.

[76] I assume in this chapter that the two general problems with respect to intentional explanations in science are adequately dealt with. These problems are, firstly, the problem of introspection as a method. This problem does not arise for Dennett as intentional attributions do not require verification through introspection. The second problem is the reductionism problem: the idea that intentional explanations should be reduced to lower level explanations. Dennett can solve this if he can show that intentional explanations provide something over and above reductive explanations, like swiftness and functional generalization (for a detailed argument, see chapter 4) and this is exactly what I aim to illustrate in this section.

7.2. The intentional stance as a method in cognitive ethology[77]

In their attempt to liberate themselves of their behaviorist chains, cognitive ethologists exactly claim, and try to prove, that intentional explanations of intelligent animal behavior possess the extra explanatory value needed to justify them.

The starting point of cognitive ethology is usually set in 1976, when Donald Griffin published his book *The Question of Animal Awareness: Evolutionary Continuity of Mental Experiences* (Griffin: 1976). In this, and later, books, he claimed that it was time for ethologists to shake off the behaviorist heritage, and to (re)start investigating the mental states of animals. Griffin, it should be emphasized, took this very literally, and argued that the *conscious* mental states of animals should be studied. Most cognitive ethologists today do not support that strong a claim; most view mental states as theoretical concepts. But a 'new' field was born, and many studies were performed that tried to map cognitive and intelligent behavior of animals, in a vocabulary that had been deemed unscientific for a long time. (To be sure, cognitive ethology is still in the stage of proving itself, and not everybody supports its research program. But it is gaining ground and support).

What kinds of processes or phenomena does cognitive ethology study? The simple answer is that cognitive ethology studies the way in which animals and other organisms ("informavores") perceive and represent their environment, come to 'know' something about it. Furthermore, it studies how these cognitive states affect their behavior – i.e. how they *selectively respond* to their environment. In principle, the domain ranges from very simple representations and actions, to really complex ones. An example of an extremely simple behavior is that of anaerobic bacteria, that detect the direction of and move towards oxygen-free water (Sterelny: 2001, 22, cf. Kornblith: 2002). But of more interest is more complex cognition-driven behavior, especially behavior that is based on some kind of representation of *another* animal (see Sterelny: 2001; Sterelny: 2003b). No cognitive ethologist that I know of accepts or finds it interesting to really include organisms such as bacteria in their domain of study, but many accept the milder claim that there is no principled distinction to be made, that bacteria

[77] In my discussion I will rely largely on the work of qualified philosophers working with cognitive ethological data, amongst which Kim Sterelny, Colin Allen and Marc Bekoff.

are *in principle* part of the domain of cognitive ethology, and that the rough structure of cognition is already present in such simple organisms.

Let me shortly contrast cognitive ethological approaches with philosophical approaches. The favorite philosophical strategy is top-down: to define central terms first, and to see what concrete empirical instances fall under it later. Colin Allen and Marc Bekoff, renowned for their putting cognitive ethological issues and insights on the philosophical agenda, make a strong plea to work the other way around. They prefer to leave the central definitions open, and fill them in as they go. They do not start with a certain fixed theory about what belief is, but aim to *investigate* that. Cognitive ethology, as such, can add new empirical material that may determine the meaning we give to attitudes over and beyond the intuitive data that philosophers usually deliver. Whether only human beings have (or can be ascribed) minds, and what a mind is, is exactly what is at stake – and not a presupposition. Cognitive ethology may force us to revise our philosophical theories.

For instance, Allen and Bekoff write very critically about the philosopher's definition of behavior-as-opposed to action. This is a hot debate in the philosophy of action, especially in the conceptualist tradition (cf. Mele: 1997 and see 2.4). Dennett's theory of action and intentionality is for a large part directed against such approaches, and cognitive ethological insights in my opinion make a good case for a more Dennettian approach towards attitudes (and for empirically informed conceptualizations more generally). The empirical reality, as described by ethologists, is that the armchair distinctions that are often at the core of philosophical action theories, especially between action and mere reflective movement, turn out to be of little value. Cognitive ethologists, Allen and Bekoff claim, are especially interested in what they call "energy-added" systems, systems that require input of energy on their part. Cognitive ethologists, then, seek to explain *why* the cost to the organism is worth paying. But even simple, hard-wired responses can be energy-adding, and are, therefore, part of the domain of cognitive ethology. Furthermore, many animals, including human beings of course, can learn to control many bodily responses. (They can resist the urge to blink with the eyes for example). So the armchair division between mere reflective behavior and real action is misleading, rather than helpful, for the cognitive ethologist (cf. Allen and Bekoff: 1997, 41-44).

> At one extreme, one might classify behavior in physical terms as changes of positions of objects with respect to some frame of reference. At the

> other extreme, one might classify it in the fully intentional terms familiar from philosophical action theory. (...) Because we do not expect there to be a way to decide this choice *a priori*, we favor a pluralistic attitude toward behavioral classification schemes (...) according to which schemes for categorizing behavior will turn out to be empirically most productive" (Allen and Bekoff: 1997, 48-47[78]).

A question that is *not* on top of the research agenda of the cognitive ethologist, is where to draw the exact line between animals that 'really' believe, represent, or know, and those that don't. Philosophers are often tempted to specify precise *definitions* of, say, belief, desire, or knowledge. Animals and other organisms know their environment in lots of different ways, some in more complex ways than others. Instead of breaking up the domain of 'believers' in strict categories, they focus on multi-dimensional characterizations of intelligent behavior. An animal may score high on one dimension, but low on another. For instance, some animals use only a limited number of perceptual clues to distinguish relevant objects and agents in the world (relatively stimulus bound), others are able to multi-track features of their environment (relatively stimulus free). Some animals have a relatively rigid behavioral repertoire (always run when you spot a predator), others are more flexible (run, hide, or make a threat – depending on the circumstances). These distinctions are not made in order to come to new fixed categories (e.g. stimulus bound = non-believer; stimulus free = believer), but are rather seen as falling along a multidimensional scale:

> the degree of interaction between internal and external factors can be conceptualized as falling along a scale. Toward one end of the scale, external stimuli predominate over internal factors and the behaviors can be considered relatively "stimulus bound"; towards the other end, internal factors predominate over external stimuli and the behaviors may be considered relatively "stimulus free". Behavioristic explanations are to be preferred at or near the end of the scale where external factors predominate over internal factors in the causation of behavioral responses. (Allen and Bekoff: 1997, 57)

Another strong example of the multidimensional approach is given by Hurley (Hurley: 2003a). She emphasizes that generality and flexibility are a matter of

[78] Last sentence quoted from page 47.

degree. But intelligence may also be domain-bound. She emphasizes that it may well be that animals are strong in applying certain schema's in one domain, and weak in another. More specifically, there could be islands of practical rationality. For instance, primates may be very good in making transitive inferences in the social domain (A dominant over B, B is dominant over C, so A is dominant over C), but fail to make such inferences in foraging contexts (tree A has more fruit than tree B, tree B has more fruit than tree C, so tree A has more fruit than tree C).

Many cognitive ethologists plea for a dimensional or gradual approach towards to investigation of behavior to action, emphasizing that the differences are typically not sharp and that it is therefore hard to say were, e.g., behavior ends and action starts[79]. But they are quick to add that, of course, relatively stimulus free behavior is much more interesting to study, and that the fruits of cognitive ethological research, and the extra value of intentionalistic explanations, is to be found there.

This is especially clear in the work of Kim Sterelny, who has mainly focused on the epistemic capacities of animals to deal with *hostile, or strategic* creatures. Biological organisms live in an *animate* world that consists of hostile and competitive creatures. This, Sterelny argues, is a key selective drive for developing higher-order representational skills. Adaptive behavior that is targeted at the inanimate world can usually rely on simple cues. "It is much more risky to depend on simple cues when interacting with rivals and enemies, since these epistemically pollute the organism's environment. The species-specific mating signals of fireflies are mimicked by predators; reliance on this simple cue is decidedly dangerous" (Sterelny: 2001, 24).

Of course, even cognitive ethologists will have to use labels, and will try to get some systematic ordering in cognitive capacities as found in the animal world. But the picture that arises is really quite different from the usual philosophical picture. The usual philosophical picture is one in which human beliefs are

[79] The focus of cognitive ethologists lies with cognition, but it is not in principle impossible to study the motivations of animals. Sterelny clears some useful ground (Sterelny: 2001, especially chapter 11), when he distinguishes so called "Nike animals" from animals that have a more complex motivational repertoire. Nike animals *just do it*, they lack a preference scale, and cannot postpone or relativize their needs, if necessary.

set off from beliefs of non-human animals, or at least one in which the domain of mind havers is clearly and firmly separated from creatures that do not have a mind (cf. Allen and Bekoff: 1997, vii-viii). Cognitive ethology extends the domain of mind-havers considerably, or, at least, is open to investigate whether and how far it should be extended. And the domain is not framed as a domain of mind-havers and non-mind-havers, but as a multi-dimensional, gradual or step-ladder continuum, on which important varieties of cognitive capacities find their place. Animals that have inflexible responses find their place on this continuum, as well as animals that have several paths to their knowledge. Animals that are quite stupid socially, but smart in chasing belong there, as well as human beings.

I have argued that attitude-terminology is methodologically justified, just in case it can be shown to have extra value in our explanations of behavior (cf. chapter 4). Cognitive ethology shows that mentalized explanations contribute to the study and explanation of animal behavior. Explanations strictly in terms of underlying mechanisms are blind to *generalities* in *kinds* of behaviors. For instance, hunting behavior can be realized in many different ways, but it is the general category of 'hunting behavior', that may interest us. (Of course, we may and will also be interested in how an animal hunts, e.g., what particular physical mechanisms trigger a hunting response, how the animal decides who to hunt etc.).

More specifically, intentionalized explanations characterize the functions of cognitive systems *differently*. Millikan, for instance, revives the classic distinction made by Ernst Mayr between proximal functions and ultimate functions. Proximal functions specify the causal mechanisms of the system in question, whereas (for Millikan) ultimate functions specify the features for which the system was selected for (evolutionary function). A good example:

> The proximal function of one of the hoverfly's visual specializations is (in [Millikan's] view) to locate small moving black dots on its retina; its remote [ultimate] function is to represent the presence of a potential mate (1990). Similarly, the proximal function of the isopod badge is to match a stored chemical sequence; the remote [ultimate] function is to admit only kin to the nest" (Sterelny: 2001, 211)

Similarly, Allen and Bekoff argue that intentional explanations are so valuable because they provide a "functional level of description of cognitive states" (Allen and Bekoff: 1997, 72, see also Millikan: 1991). Functional classifications allow us to group together different kinds of behaviors, with different underlying mechanisms.

In general, cognitive ethologists justify their explanations as higher-order program explanations (as opposed to process explanations) *a la* Pettit and Jackson, program explanations that capture the 'design function' or 'ultimate functions' of cognitive states. Such program explanations do not conflict with reductive (process) explanations (cf. 4.2). (Dennett may add that intentional explanations also provide some practical advantages, such as swiftness.)

Although there is no principled reason to refrain from explaining the behavior of simpler organisms in terms of cognition, the most interesting questions lie with the 'smarter' animals that are capable of multi-tracking their environment. Cognitive ethologists run the risk of overattribution. We have seen that intentional explanations are justified exactly because they explain behavior from a different, more general and programmatic, perspective. At the same time, they do not want to overattribute intentionality to animals. A standard example is the behavior of the piping plover, a small bird, that seems to be capable of some clever deception (see Ristau: 1991). When a predator approaches its nest, it will try to direct the attention of the predator towards itself, and away from the chicks. It will feign to have a broken wing, making an attractive prey for the predator, lure it away from the nest, and fly away when the predator makes its attack. Is this bird really a clever bird that succeeds in deceiving the predator? Or is its behavior merely hard-wired, triggered at every instance (even when it is not really clever) a predator approaches the nest.

Cognitive ethologists often invoke the *principle of parsimony* (also called 'Morgan's canon') in such cases: if you can explain the behavior strictly at a lower level, you should refrain from explanation at the higher level. But it is exactly at stake whether lower level explanations *can* substitute the higher level explanations because the claim is that intentional explanations (or, for that matter, functional explanations), explain something *else* than lower level explanations (see chapter 4). As we saw, even the behavior of very simple animals can be framed at the programmatic, functional level.

Consider this example from Enç. Some fruit flies have a sensor that help them seek humid spots (they need humidity in order to survive). The sensor, however, does not directly seek humidity, but responds to intensity of light in the environment, steering them to darker areas, that are, *in the fruit flies' environ-*

ment, usually humid. At the causal level, the sensor can best be described as a device that detects a certain intensity of light. But at the functional level, Enç argues, the device is best thought of as a humidity-seeking device (Enç: 1982, 168). According to Enc, such functional explanations are very valuable for understanding psychological mechanisms. Not only do they offer a richer account because they do not only tell us something about the causal mechanisms involved, but also about the reasons why these mechanisms are in place. (Enç: 1982, 169). Should we, in this case, use the principle of parsimony, or shall we allow for both types of explanations? I think the second option is best: allow for both types of explanations. Only then can be done full justice to the idea that intentional explanations capture *something else* than lower level explanations, and according to the non-reductionist account of explanation, such higher order explanations are acceptable exactly in such cases.

There is an important qualification to be made at this point. First of all, *contra* to what I have suggested above, many cognitive ethologists are hesitant to use attitude talk in their scientific work. For instance, Sterelny says:

> [attitude talk] does not even carve *our* sensing and control mechanisms at the joint, it would be a miracle if it were well-suited for describing those of nonhuman agents (Sterelny: 2003a, 259)

And Peter Godfrey-Smith writes:

> If we think of folk psychology as a socially-evolved interpretative tool that functions to help us deal with a specific set of social tasks, then when it is used to describe non-human animals *it is far from its domain of normal use* (Godfrey-Smith: 2003, 267)

Both quotations are from a discussion with Hurley (Hurley: 2003b), who defends the application of attitude talk in cognitive ethology by means of an interpretationist approach. But this is a matter of terminological misunderstanding. Sterelny and Godfrey Smith have a specific meaning of attitude terms in mind. They see it as something specifically human, terminology that human beings use to explain and predict *each others* behavior. In addition, they believe that the content of beliefs and desires has to be determinate – and that this determinacy can never be reached in the animal domain. But it is not at all clear whether beliefs should be seen as having determinate content, and it is not at all clear that human beliefs have such a determinate content. This would count against a science of beliefs at all, and not only against a science of beliefs of animals. At the same time, they frequently use terms like cognition and motivation, which have, I think, the same meaning as terms like belief and desire taken liberally. As I see it, Sterelny and Godfrey Smith are skeptical about using the standard (realistic) meaning of folk psychology for animals, for the same reason that philosophers like Dennett (and Hurley) reject using such folk psychological notions for human psychology. The vocabulary is, unrevised, just not a good vocabulary. But they may be happy to embrace it if the basic terms – 'belief', 'desire', etc. - are reconceptualized, as Dennett suggests (cf. Allen and Bekoff: 1997, 2).

Let me put it differently. We may claim, as Sterelny and Godfrey-Smith do, that folk psychology is a flawed theory, and should therefore not be applied to human beings, let alone to animals. Or we may claim, conversely, that folk psychology is a flawed theory, that we should therefore revise the terminology, so that we can apply it to human beings and to animals (thesis of attitude generosity). The glass is half-empty, or half-full. I believe that the second strategy is the strategy that is *in fact* taken by cognitive ethologists. The categories they use are revised attitude-terms, and the explanatory framework has interpretationist characteristics (e.g. holistic determination of the content of cognitive states). They may merely be hesitant to actually *use* attitude terms, if only because it may give the wrong associations to readers, and in order to steer away from different discussions that can only harm their cause.

7.3. Dennett and cognitive ethology

Cognitive ethologists have shown the extra value of attitude explanations in the behavior of a wide range of animals, including relatively simple organisms. But remarkably, Dennett seems to be hesitant to embrace such explanations in cognitive ethology and to use them to support his theory of the intentional stance. This is remarkable given Dennett's own principle of generosity. Cognitive ethology lends support to a methodological justification of the principle of attitude generosity, but Dennett kindly seems to *reject* it. This is particularly remarkable, because *if* the principle of generosity is to be justified, one would expect it to be a methodological justification, much like the one cognitive ethology is offering. In other words, if *cognitive ethology* is not an adequate example of the principle of generosity at work, what is? Is their offer an offer Dennett can refuse? I do not think he should.

Dennett has an extremely liberal account of attitudes, as I captured in the principle of attitude generosity (chapter 5). Whenever we find it useful to attribute intentions to a creature, we should and may do so:

> Any object – or as I shall say, any *system* - whose behavior is well predicted by this [intentional] strategy is in the fullest sense of the word a believer (Dennett: 1987f, 15)

And:

> The claim that *in principle* a lowest-order story can always be told of any animal behavior (...) is no longer interesting. (...) Today we are interested in what gains in perspicuity, in predictive power, in generalization, might accrue if we adopt a higher-level hypothesis that takes a risky step into intentional characterization (Dennett: 1987b, 247)

But when it comes to science, and the science of cognitive states in animals in particular, Dennett sometimes seems to hold a much more conservative position, tending towards a strict application of the principle of parsimony. This impression rises not because he explicitly says so, but rather from the way he discusses animal beliefs in scientific contexts. First of all, his own scientific examples of using the intentional stance in cognitive ethology are examples of vervet monkeys that show some kind of *higher order* intentionality, beliefs about beliefs for instance (Dennett: 1987b). Secondly, his knock-down arguments against behaviorism in ethology rely, again, on predictive and explanatory success of intentional explanations regarding *higher order intentional behavior*, for instance, deception (Dennett: 1978d). Indeed, many cognitive ethologists (wrongly) regard Dennett as an eliminativist about attitudes (see, e.g. Allen and Bekoff: 1997, chapter 5, see also Seyfarth and Cheney: 2002).

Dennett tries to relieve this tension with his notion of 'free floating rationales'. Even though we probably cannot credit the piping plover with an intrinsic system of intended deception, but there is still a sense in which *there is a rationale* behind its behavior.

> The deceptive rationale is there all the same, and to say it is *there* is to say that there is a domain within which it is *predictive* and, hence, explanatory (Dennett: 1987b, 259)

There is a reason why the bird acts like it does - *it is no accident* - and it *is* a case of deception, even though it is hard-wired. The rationale, in this case, lies in nature, more specifically, in the process of natural selection. We can still use the intentional stance to trace such rationales, but now we do not target it specifically to the bird itself, but we redirect it to nature, or rather, the process of natural selection ("Mother Nature"). The methodology of the intentional stance, then, is applied as a tool to find the rationales that have evolved in nature, the *non-accidental*, designed features of animals and other organisms. Here, again, it is *predictivity* (and explanatory power, perhaps) that legitimizes the use of the intentional stance in science.

Still, Dennett's move to the idea of a free floating rationale seems too cautious. We have just seen that recent work in the field of cognitive ethology sufficiently justifies the generous scientific use of intentional explanations (ergo, the intentional stance), and that it justifies it exactly in the domain that matters: epistemology. Dennett has a great case for the (generous) intentional stance here, but he backs out. As we shall see, he makes a similar move with the design stance, that is set up equally liberally, but whenever things threaten to get serious (i.e. when it comes to 'real' science) Dennett exercises restraint. There is a clear tension between, one the one hand, Dennett's wish to be generous with respect to the application of the stances and his wanting to justify the daily stances, and, on the other hand, his ambition to hold on to the (classical) standards of scientific explanation.

I believe that the tension described above, relates to a certain tension between and within Dennett's pragmatism and naturalism. The intentional stance, we might say, is a clear product of Dennett's pragmatic views, whereas the design stance represents his naturalistic consciousness (cf. 5.5), Dennett, as argued at length (3.5), has a great ambition to construct a naturalistic philosophy. What makes his approach original and attractive, is the pragmatic twist he gives to his naturalism, claiming that scientific approaches are just very practical (because scientific experiments are repeatable, predictable, falsifiable, etc.). An original combination, but here we find a clear case where Dennett has to carefully maneuver between the 'rules' of pragmatism (just show that it works) on the one hand and the other 'rules' of science (such as Morgan's canon and the ambition to show *why* things work) on the other. Like the egoist that sees profit in doing altruistic deeds, 'investing' in scientific principles may be very practical. But there is always the question of when to invest and when to cash out.

We have already seen how the tension between pragmatism and naturalism, between a generous intentional stance and a strict design stance, pull Dennett in different directions (5.5). Dennett's hesitance in the cognitive ethological domain, is, I think, an instance of the very same problem. In this case, however, the tension is not necessary. Cognitive ethology provides a perfectly natural example of where pragmatic considerations and naturalistic standards go nicely together.

7.4. The design stance as method in archaeology[80]

Does Dennett's account of the design stance, the optimality account, provide a good scientific method for ascribing functions to artifacts? Dennett does not have to worry about reductionist concerns, they should be out of our way (see especially chapter 4). The only thing that is required now is to show the virtue of the optimality account as a method in science. In order to do that, the optimality account will have to prove its worth in competition with its main rival: the intentionalistic account.

Many scientific fields use a notion of function, especially in the humanities and social sciences. But artifact function is not often studied. I believe the natural field to critically test the design stance is archaeology, in fact a domain Dennett often refers to. The archaeologist ascribes functions to artifacts – tries to find out the function of an artifact – and tries to do that in a scientifically respectable way.

The domain of archaeology is interesting to discuss for another reason. It is interesting to compare the notion of function in archaeology with the notion of function in evolutionary biology. Most theories of artifact functions are framed in comparison with theories about biological function. Some theorists, especially defenders of the strong intentionality claim, argue that the two notions are different (e.g. McLaughlin: 2001, Vermaas and Houkes: 2003, others claim that they are basically the same (Preston: 1998b, Millikan: 1993). Remarkably, however, all these theories tend to compare the *methodological* notion of biological function with the everyday, *intuitive* notion of artifact function. 'Biological function' is rarely conceptualized in everyday terms; rather, the philosophy of

[80] This paragraph is based on joint work with Krist Vaesen (Vaesen, K. and Amerongen, M. v.: forthcoming).

function conceptualizes some kind of scientific notion of biological function. And 'artifact function' is rarely conceptualized in scientific terms; rather, function theories conceptualize the way we use the term artifact function in everyday life. Constructing a scientific notion of artifact function would help to make a proper comparison.

But most importantly, archaeology provides a critical case to test Dennett's account of the design stance. As we shall see, there is very little ground for intentionalistic reasoning in archaeology. The first part of this section will be devoted to showing exactly that intentionalistic accounts of function are in a great disadvantage in comparison with non-intentionalistic accounts of function.

I have to be careful about what I claim about archaeology and its methodology. This is not an archaeological thesis, I am not an archaeologist, and most of all: there is a lot of controversy over what the proper methodology is for an archaeologist. Teleological explanations are, in archaeology, as suspicious as they are in evolutionary biology, psychology, and ethology. The line of argument of the reductionist non-teleologicians is quite similar to those of behaviorists in these other sciences ('science has no place for intentionalistic and normative talk'). Fortunately, I only have to deal with those methodologies in archaeology that say that teleological reasoning *is* admitted in the field. Dennett's optimality account of function is a teleological theory, and if it is supposed to show its worth in archeology, it must do it as a teleological theory. Moreover, if these behaviorists are right, the optimality account is as wrong as the intentionalistic account[81].

Why should an archeaologist want to use functional explanations? Salmon says about them:

> Functional explanations are as important to the archaeologist as to the evolutionary biologist, and they are used not only to explain evolution, but also to explain why prehistoric people settled at particular sites, why certain items occur in these sites, why tools or facilities have a particular form, or why certain processes (...) occurred. However, many archaeologists would hesitate to call these explanations 'functional'. 'Systems explanations' is a

[81] Note that the causal role notion of function ("Cummins-function") is not an alternative as it is not a teleological notion of function. Etiological notions of function may be an alternative and will be shortly discussed at the end of the chapter.

> currently popular euphemism (...) these 'systems' explanations have all the essential features of functional explanations (Salmon: 1982: 87)

I will focus on attempts to fruitfully use teleological explanations in archaeology, and on arguments from archaeologists to do so. In this way I can tentatively answer the question: *for as far as* teleological explanations are justified and fruitful in archaeology, what kind(s) of notion(s) of function would they require? And: can such a notion be devoid of any reference to intentions?

7.4.1. Optimality reasoning in archaeology

Archaeology *should* speak greatly in favor of the optimality account, in any case versus intentionalistic accounts. In the very first place, of course, because information about intentions is hardly available. There are few written sources about these times which makes it very hard to find out what was on the early hominids' mind; on the contrary, the material remains that are left by our prehistoric ancestors are used to *reconstrue* their beliefs, wants and intentions. In that sense, archaeology takes the demanded direction of the optimalitist: from material facts to intent. The optimality account, then, has the epistemological upper hand.

The lack of information about the mental life of the early hominids has pressed archaeologists hard to determine functions of artifacts without being able to refer to intentions. And indeed, often it is not necessary to refer to intentions in order to construe plausible hypotheses about functions of artifacts. Take, for instance, the following rule of thumb for the archaeologist:

> "The more severe the limitation on the form of an object that the suspected function imposes, the more reliable is the ascription of that function (...). For some objects, such as grinding stones, there are very few forms that are compatible with reasonably efficient performance of the function" (Salmon: 1982, 59)

The principle is relevantly similar to Dennett's optimality principle that it would count against something being a cherry pitter, if it would be a demonstrably inferior cherry pitter. And indeed, we may expect that there is a certain match between the form and features of an artifact, and the function we ascribe to it – especially if it is found in large numbers.

Dennett refers to the experimental research by William Calvin on the function of the much debated Acheulean hand-axe in archaeology in order to defend his optimality account against intentionalistic approaches. Let me shortly discuss a few (anti-intentionalistic) claims about this hand axe. They all seem to speak in favor of Dennett, but as I argue later, there a few important qualifications to be made.

The famous and mysterious Acheulean 'hand-axe' is a stone item that has been found in great numbers at many archaeological places. It is thought to have been made around 1,5 million years ago (cf. Gibson and Ingold 1994). Their massive presence at archaeological sites and the similarities in form strongly suggest that the item *had* a function, and its form (especially its literally "handy" size and the pointy head) suggests that it was a hand axe. But as a hand-axe it would be quite useless: the edges are so sharp that hominids using it would cut themselves. As such, the hand-axe has been an object of massive archaeological speculation. Was it indeed a hand-axe? For what purpose was it made? What kind of information would we need to be able to answer such questions? Can we answer such questions at all?

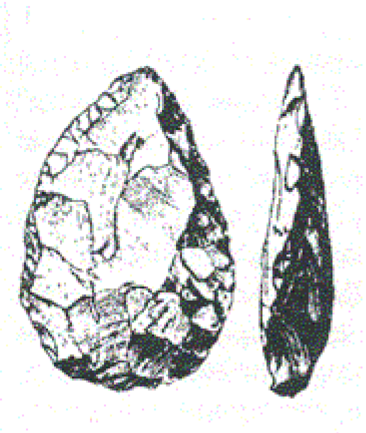


Figure 7: The Acheulean 'hand-axe'[82]

William Calvin[83] has hypothesized that the hand-axe was not really a hand-axe, but rather a throwing stone. By means of performance studies he found that the stone could be thrown as a discus at a herd of large mammals (probably cows).

[82] Drawn from http://williamcalvin.com/BrainForAllSeasons/Olor.htm

[83] http://williamcalvin.com/1990s/1993unitary.htm

This would both explain the form and the sharp edges of the biface. The form is perfect for throwing the stone as a discus. The sharp edges insure an incision pain, causing the animal to act on reflex and to fall down, not being able to remain in balance.

Calvin's hypothesis is indeed a good example of a reasoning by appeal to optimality: assuming that the biface *had* to fulfill a reasonable function and that the parts, form and shape of the artifact are like they are for a *reason*, he constructed and positively tested a functional hypothesis whereby the biface indeed fulfils an optimal function, and all features of the stone can be functionally explained.

A second hypothesis is defended by Davidson and Noble. In this case, there is no direct appeal to optimality, but they make a strong case against strong intentionalism, or, more specifically, against blueprint-conceptions of artifact construction. The seemingly careful design of the artifact, one might think, must point at some intention (in the strict sense) to create a specific form, a blueprint in the mind of the makers. Such an inference is of course especially interesting for those who want to prove that the early hominids were indeed capable of forming blueprints in their minds, as well as able to plan and execute a certain procedure to actually realize this blueprint. Davidson and Noble, amongst others, aim to show exactly that this conclusion is much too quick. They argue that bifaces are residual cores left after successive removals of flakes. Hominids may have used the cores for some function, but it was primarily the flakes they were after. The *flakes* were meant to be used as knives, not the cores. Thinking that the *cores* are the real artifacts, not the flakes, is in their terminology making *the finished artifact fallacy:*

> The fallacy is the belief that the final form of flaked artifacts as found by archaeologists was the intended shape of the 'tool' (...)
>
> Part of the finished artifact fallacy is that categories recognized by archaeologists were forms created, successfully, by self-conscious prior intent (...) Many of these 'steps' derive from the self-conscious decision making necessary for the experimental replication of desired end products. But that does not imply similar decision making of self-consciousness by the prehistoric knappers. Once it is conceded that some of the products archaeologists observe are 'partly completed' or 'failures', the whole question opens up. None of the classification schemes admits the possibility of 'partly completed' or 'failed' products, yet they are an evident outcome of

> experimental replication. It seems less certain, then, that we can make unqualified judgments that there were *internal* end products (Davidson and Noble: 1994: 367)

Davidson and Noble argue that archaeologists often are easily biased by supposing that the form of an artifact was envisioned by the hominid toolmakers as mental plans or blueprints, and subsequently created according to that image. This bias leads us to think that the cores, and not the flakes, were the intended products of the hominids. But this need not be the case. It is very well possible that the similarity in form of the bifaces that are found, are the result of a certain way of striking the flakes from the core. In action-theoretical terms: the form of the core is not 'caused' by the intention of the designer to create that form, but by the intention to create a certain flake – if we can speak of intentions at all (Davidson and Noble 1994).

> …a single flake may be repeatedly used and reused, until it is eventually discarded. To call this discarded form a 'finished artifact' is like saying that the 'finished pencil' is one that – through repeated use and resharpening-has been reduced to a stub (described in Gibson and Ingold 1994: 340)

Indeed, experimental studies by Shick and Toth also show that the flakes and the corresponding cores could have been produced without any planning ahead. Stones can be stroked only in a limited number of ways, and each way produces a typical kind of core – the kinds that are indeed found at the sites. The similarities in form of the Acheulean 'hand-axes' could then be created by creatures without a fully developed mind - and should therefore be explained without appeal to any prior represented intentions (Schick and Toth: 1993):

> Oldowan tools are technologically quite simple, and many or most of the so-called 'core-tools' (choppers, discoids, polyhedrons, heavy duty scrapers, etc) appear simply to be discarded cores manufactured in the process of flake production. These core forms are not necessarily tools, nor do they necessarily correspond to 'mental templates' held by early tool-making hominids. The final morphology of these core forms may be determined largely by the size, shape, and raw material of the rock used (Toth and Shick 1994, 349)

There are more reasons to reject an appeal to strong, represented or "literal" intentions in archaeology. Cognitive archaeology, a branch in archaeology that tries to infer something about cognitive capacities of the early hominids on the basis of found artifacts, is under a lot of strain. It is extremely difficult to prove that hominids were, say, able to project a designed form onto a to-be-created artifact, on the basis of found artifacts (cf. Salmon: 1982, and more recently, Renfrew and Zubrow: 2000, on cognitive archaeology). Whatever the results of these debates will be, it should be clear that there is a strong case to be made that we cannot *start* from hominid intentions when reasoning about the possible function of found artifacts. Whether they even *had* 'intentions' in the strict intentionalist sense, is exactly what is at stake in these discussions.

7.5. Optimality and intentionality

Thus, research on the hand axe gives rather different, but plausible, functional explanations of the 'hand-axe', but all in which appeal to (strict) intentions is either unnecessary (Calvin) or explicitly rejected (Davidson and Noble). Completed with the little information we have about the mental life of our hominid ancestors, archaeology seems to be a settled case for Dennett. There are, however, some notable qualifications to be made that should warn us against drawing too easily the conclusion that Dennett's optimality account wins the battle against intentionalism in the domain of archaeology.

Let me start by taking a closer look at Davidson's and Noble's 'blueprint argument'. Suppose that the early hominids did not have 'intentions' in the intentional realist sense, that they were unable to plan ahead, and that they were not capable to project designs in their heads. This appears to speak against intentionalistic accounts of artifact function, for the simple reason that there would then exist technical artifacts that were not created with a clear designer intention. The intentionalist might in that case of course just bite the bullet and claim that hand axes and other artifacts created by these early humans are not part of the set of functional objects, but clearly it would be a defeat to exclude the first man-made artifacts from the domain of technical artifacts.

Unfortunately for Dennett, the blueprint argument does not settle the battle between intentionalism and optimalitism. Let's start with the blueprint argument. The blueprint argument is strong *only* against those theories of artifacts that say that only consciously represented, full blown intentions are capable of transferring functions to artifacts, but it has little power against intentionalistic theories of artifact function that take a more liberal approach towards attitudes (cf. the distinction between intentionalistic accounts of function I made at the end of 5.8.2).

As a matter of fact, were it not for the fact that Dennett must refrain from using the intentional stance when interpreting design (5.6) , Dennett's generous intentional stance *might* have been the perfect tool to say something about our ancestor's 'thoughts' behind their technological inventions. For if we can talk in a scientifically respectable way about the intentions of vervet monkeys, of killer spiders and of thermostats, we should certainly be able to say something about the intentions of our own ancestors (as designers)! The very strength of the intentional stance is that it shows that we can often be agnostic about the exact developmental status of the mind that belongs with the behavior that we try to explain, human beings and other animals alike. That the results of such intentional reasoning may be indeterminate and sometimes unreliable is just part of the game and is not principally different from reasoning about intentions of other creatures. So if the early hominids were relatively 'simple minded', this is not a reason per se to refrain from intentional reasoning. On the contrary, they might be a welcome bridge between human technology and technology of other animals, like bird nests and bee hives,

In other words, the blueprint argument forms a challenge to strong intentionalism, but it is neutral towards the battle between optimality and other forms of intentionalism. But this also means that the blueprint argument does not speak *against* optimality. In order for my argument to work, it is not enough to show that there are other theories available. For the critical case strategy to work, I need to make plausible that optimality is the wrong choice, or at least that these other theories are *better*. But if I can make plausible that design interpretations are better (more useful, more informative, more complete) if we do take into account designer intentions, I *do* have a case against optimality.

So let us take a closer look at the ascription of optimal functions to the hand-axe. Calvin's experimental research showed that reasoning in terms of optimality is indeed an informative strategy, but I believe that such reasoning presupposes at the very least some vague hypotheses about the intentional background against which such ascriptions are made. Suppose, for instance, that the Acheuleans were strictly vegetarian, implying that they did not *need* or *intend* to kill mammals at all. In that case, the production of hunting gear would be irrational, making the killer discus hypothesis much less convincing. As Krist Vaesen (forthcoming) nicely put it: the hypothesis that the artifact is a killer discus is counted against by the fact that it would have been a demonstratively *useless* artifact. Calvin's experiments have retrieved one of the objects *possible* capacities, but the reasonableness of this suggestion is dependent on a set of implicit assumptions about the Acheulean intentional life. If we want to know whether a possible optimal function of an object was in fact the function for which the object was used, it seems that we at least have to know something, or assume something, about what the Acheulean hominids wanted and needed, and what they could know and do. And these, of course, are all intentional terms[84].

That function ascriptions may be seriously misguided when we neglect intentional contexts can also be illustrated by means of another, more recent, example: the Babylonian Battery. Near Baghdad many so-called Babylonian batteries have been found, objects belonging to the Parthian (between 250 BC and 224 AD) or Sassanian era (224-640 AD). Each so-called battery is a 15cm vessel which contains a cylinder of sheet copper, capped at the bottom, in turn covering and protecting an iron rod, which shows signs of acid corrosion (see Figure 8).

In 1938 Wilhelm König, director of the National Museum of Iraq, published a paper suggesting that the artifact may have been a galvanic cell, a kind of primitive battery (König: 1938 see also Dubpernell: 1978). And that would indeed be a spectacular discovery if it were true. It would imply that electrical current had been used by the ancients and was only rediscovered by Galvani and Volta, some 1,800 years later.

König's hypothesis was tested by Jansen et al. (Jansen, Fickenfrerichs et al: 1993). They concluded that, when fueled with a solution of benzoquinone, a substance occurring naturally in the secretions of some beetles, the object could indeed produce a certain voltage (+/-0.87V); in other words, physically the artifact can fulfill the function of a battery. The question—to Dennett— then is: is this sufficient to call the design optimal, and thus to conclude the thing to be a galvanic cell?

[84] Note that there is a serious threat of circularity here. For how do we determine the content of such beliefs and needs? An obvious and actual choice would be to look at what kinds of tools they used: using hand-axes suggests that they killed animals. We may need further evidence to try to settle the issue (we may, for instance, look at the way their teeth are formed). But then the case may still be indeterminate. I take it that this hermeneutic circle is part of the interpretative game and even more reason to include inferences about intentions in it.

A "yes" would be—and has been—wholeheartedly embraced by proponents of the paranormal. Some of them (see for instance Ortiz de Montellano and : 1991; Von Daniken: 1993) find in the "power source" thesis evidence for a technologically advanced (extraterrestrial) civilization in remote antiquity. Most historians, however, are fairly skeptical (see for instance Eggert: 1996), since the alleged galvanic cells are contemporary with the growth and height of the Roman Empire. The latter is a fairly well-documented era, hardly a period thus in which such a civilization would have gone unrecorded, particularly when the Parthian Empire was Rome's principal enemy in the east. Furthermore, what kind of appliances would need the energy supplied by these cells? Until now, no Parthian electronic devices have been found, so at the face of it, there was no Parthian desire—and hence no intention—to produce electricity whatsoever.

What do we learn from the case of the Babylonian battery? The hypothesis that the Babylonian vessel was used as a storage device for sacred scrolls, which were wrapped around the iron rod—as, among others, Paszthory (1989) suggests—seems more plausible and is more accepted than the "power source" hypothesis, but not unequivocally in the light of optimality. Indeed, the vessels are good at storing parchment or papyrus, but why would they need an iron rod and a copper cylinder (and an asphalt seat at the bottom and an asphalt stopper on top)? In engineering terms, the artifact is over-designed; and thus it would be in conflict with Dennett's optimality principle which states that every artifact's component should have a *raison d' être*. Besides, if one thinks over-design is unproblematic, why not just claim the artifact to be a container full stop, or a container of air? Because without the restriction of over-design, such hypothesis would be just as reasonable as the "power source" hypothesis. The case of the Babylonian battery shows that over-design is sometimes only explained by reference to an intentional context.

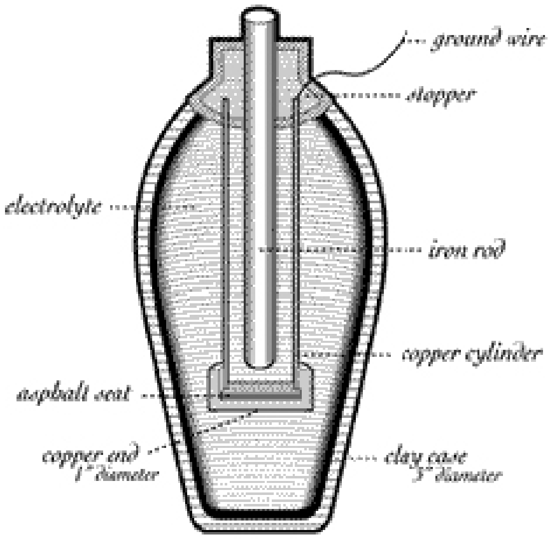


Figure 8: The Babylonian 'battery"[85]

Paszthory reasons in terms of intent, not mere optimality. He explains iron and copper to have a symbolic meaning, playing a role in ancient alchemy. According to the Parthians, it might well have been that the use of these metals was meant to please the gods, an attempt to protect their sacred documents against divine terror. As such, Paszthory hypothesizes about what Parthians could and couldn't have meant the artifact to be for—inferring from knowledge about their *beliefs* and *desires*—to adjust his interpretation accordingly. The example illustrates that, pace Dennett, a methodologically sound interpretation of an artifact may well involve an interpretation of intentions, beliefs and desires.

Successful experiments in experimental archaeology, as Eggert (1996) remarks, can only show a supposed ancient technique to be possible, but never its actual application. For instance, when Thor Heyerdahl crossed the Atlantic in an Egyptian boat, he only showed that it was possible, in principle, for the Egyptians to have done likewise. But to accept the claim that they indeed did, one would need archaeological evidence from America. Similarly, one should be wary of interpreting the experimental evidence about the Babylonian vessels as decisive. Even if the vessels can generate current—perhaps even optimally—this

[85] The example is from a draft of the dissertation of Krist Vaesen, and I thank him for permission to re-use it here. The picture can be found at several places on the web, e.g. http://www.akri.org/museum/images/bagbat.gif.

doesn't prove that they once did. In contrast, the example shows to what kinds of anomalies optimality conditions can lead, if not constrained by some kind of intentional reading; we sincerely do not believe extraterrestrial visitors once learnt the Parthians to produce energy. And if they did, they were remarkably parsimonious; being visitors with space travel capabilities, why didn't they show the Parthians less primitive ways of producing energy?

The main point I want to make is that a theory of function attribution is simply better if we allow *reference to intentional contexts*. Function attributions may be seriously misguided if reference to beliefs and desires of users *and* makers is avoided.

To be sure, Dennett certainly does not deny that context is important, but in the case of human-made artifacts, this context will often be intentional, as the case of the Babylonian Battery showed. 'Optimal' is a term that is always used relative to certain constraints. "Given the lack of funds, this design would be the optimal choice". Or "given that they had to come up with a solution soon, that design was the optimal choice". Some of these constraints are clearly intentional, like the current state of knowledge, goals, and wants. The Babylonian battery isn't a real battery, because human beings *didn't know* it could have that function, or perhaps did not even *want* to use it for that purpose (religion may have been more important than economy). Beth Preston has argued that a non-intentional and non-contextual approach is not going to provide satisfying answers to some of our most pressing questions, nor is it going to cut technological reality at its joints. Again, the case of the Acheulean hand-axe helps to show a point. Preston says:

> In the absence of any independent non-individualistic information about the normal use history, we may not be able to tell the differences between a tool and a non-tool, even when we can reconstruct the causal history of the individual object. (...) In the case of the Acheulean 'hand axes', for instance, you can hold the object and the process of its production constant, and then, by varying the local practices of the makers and the users, the 'hand axe' can be made to shift back and forth from being a designed and manufactured meat clever to being the discarded by-product of the process of manufacturing kill-blades (Preston 1998, 520-521)

So, even if we are able to reconstruct exactly in what way the artifact was created, we need at least *some* information about the intentional context in which it was developed, in order to construct plausible hypotheses about its function. Mere knowledge of how and when a certain artifact was created, does not give us an answer to the question what the artifact was created *for*. This context can be partly recovered in a non-intentional way, for instance by showing that a certain tool reasonably could not have been used for a certain task, or by showing that it could have been. William Calvin's experimental studies are an example of that. But it seems hard to reconstruct the context in a purely non-intentional way and without such reference, our hypotheses may be unnecessarily misguided.

Let me take another example. Dennett writes about the mysterious Antikythera mechanism. The function of this ancient artifact is still debated by historians. Was it a clock, a calculator, or even a planetarium? Dennett sides with those who claim that it was an orrery or planetarium, because:

> the proof of that is that it would be a *good* orrery. That is, calculations of the periods of rotation of its wheels led to an interpretation that would have made it an accurate (Ptolemaic) representation of what was then known about the motions of the planet (Dennett: 1990, 184)

Fine, but to what extent does this example speak for the optimality account? The Antikythera mechanism was only an optimal orrery *given* the beliefs about the universe of the people who made it at the time. From our perspective, it is probably a very bad planetarium, as we are able to create a much more detailed and better image of the Galaxy, and thus for us it is mainly of historical interest. To put it differently, the logic of the optimality account demands that we reason from optimality to intent (or more broadly, to intentional states), so it is only *after* showing that the orrery is a good orrery, that we may conclude that *apparently* their knowledge of the universe was limited and thus *apparently* intended it to be an orrery.

Perhaps suprisingly, Dennett himself makes this very point when he claims that:

> [w]hat appears far from optimal on one set of constraints *may* be seen to be optimal on a larger set. The ungainly jury-rig under which the dismasted sailboat limps back to port may look like a mediocre design for a sailboat until we reflect that given the conditions and available materials, what we are seeing may just be the best possible design. Of course it also may not be. Perhaps the sailors didn't know any better, or got rattled, and settled for making a distinctly inferior rig. But what if we allow for such sailor ignorance as a boundary condition? "Given their ignorance of the

> fine points of aerodynamics, this is probably the best solution *they* could have recognized." When do we – or must we – stop adding conditions? There is no principled limit that I can see, but I do not think this is a *vicious* regress, because it typically stabilizes and stops after a few moves, and for however long it continues, the discoveries it provokes are potentially illuminating. (Dennett: 1987b, 264)

I fully agree that such reasoning is illuminating and I think that this is in fact the proper way to interpret human-made artifacts, but unfortunately for Dennett, this direction of reasoning is exactly what is *not* allowed in Dennett's optimality account - as argued for in chapter 5. Not only does Dennett suddenly grant that our hypothetical designer to be less than ideal (to say the least!), but he has designer(s) intentions explain design – and this is a move he just cannot make on pain of circularity. So either I misunderstood Dennett after all, or he should withdraw the above quotation. I stick to my guns. If Dennett wants to accommodate intentional conditions after all (as he suggests in the quotation), he shall have to show how this can be done in a consistent way. Until then, I submit, Dennett's has a less-than ideal methodology of artifact interpretation with little room to improve it.

7.6. The failure of the optimality account as method

The previous paragraph should have made clear that even in an intention-poor field like archaeology, the optimality account is quite limited. Dennett has not provided us with a lot of instructions to construct a full blown methodology, let alone a clear idea how this optimality account should be fruitfully applied. If we apply it to a field like archaeology, as I have done, the performance of the design stance is especially low compared to Dennett's methodology of the intentional stance (e.g. cognitive ethology).

Granted, from the archaeological point of view, it cannot be that we start our theory of function with the ('strict') intentions of a designer, a blueprint of a design, or any specific communication of the purpose of the artifact to the projected user. It is very unlikely that our hominid ancestors did have the cognitive capacities for such reasoning; it is likely that the artifacts they created were in some sense selected because they fulfilled a certain function – hence, the notion of artifact function does not necessarily start off from strict designer intentions (or user intentions for that matter, see p. 118 for the distinctions between forms of intentionalism). Some archaeologists therefore prefer to speak in terms of tools, not artifacts (Schick and Toth, see also Dipert: 1993), but many use the terms interchangeably, and emphasize that no strict division can be made or should be made.

But that we should be extremely careful using strict or extreme intentions in archaeology, does not at all speak directly in favor of the optimality account. Tracing functions of ancient artifacts requires at least a lot of knowledge about the intentional context, *especially* if we are after their optimal functions. This is especially so when we are dealing with symbolic functions (like the Babylonian battery).

It is clear that optimality considerations are of enormous importance for generating hypotheses and theories. But the plausibility of such hypotheses, and their experimental validation, goes beyond the simple rule of thumb that we should in principle expect artifacts to fulfill or to have fulfilled an optimal function. There are many examples where optimality considerations alone would have let us astray very quickly – towards grave overattribution of the capacities of our hominid ancestors, rather than an underestimation. Even if the intentions of early artifact designers and users are indeterminate – as said, a reason for Dennett to discard intentionalistic approaches – this does not mean that it is unproductive to hypothesize about them; indeterminacy is simply part of the hermeneutical game.

I conclude that if Dennett wants his account to answer questions about what a certain artifact was historically designed for (the historical question), he will have to accommodate intentional context in his optimality account, which means that the optimality account turns into an intentionalistic account - which he does not want.

At this point Dennett may respond that he is not at all interested in the historical or actual function of artifacts and that I am barking at the wrong tree. After all, Dennett explicitly writes that what designers intended an artifact to be says nothing about the use others might find in it.

> The inventor is just another user, only circumstantially and defeasibly privileged in his knowledge of the functions and uses of his devise. If others can find better uses for it, his intentions, clearheaded or muddled, are of *mere historical interest* ... (Dennett: 1990, 186 & 194; quoted earlier in chapter 5).

The point is well taken, but then I wonder why Dennett makes so much use of these historical examples to explain his optimality account (think of the Anthykera mechanism, about the examples in archaeology, etc.). Furthermore, this would imply that the relevant questions about functions in historical studies – What was this object created for and why? - are unanswerable by the optimality account. But assuming that the optimality account is not supposed to say anything about the original intended functions, the immediate question that rises is: What scientific or methodological use might the optimality account have to offer then?

7.6.1. The (lack of) predictive value of the optimality account

In his quest against intentionalistic accounts of function, Dennett claims that his own optimality account has better predictive credentials. The optimality account is supposed to provide reliable predictive function ascriptions. As such, it has the best credentials for being turned into a practical (scientific) method. Predictivity thus might compensate for the disappointing score until now on the methodological scale, so it is a relevant argument to take into consideration.

Recapitulating, Dennett claims that the design stance may help us predict (artifact) behavior when we know what the function of an item is (5.2). Fair enough, but does the (optimality account of the) design stance also have a predictive upper hand when our task is to find out what the function of the item is? Dennett certainly suggests it when he writes that:

> The inventor is just another user, only circumstantially and defeasibly privileged in his knowledge of the functions and uses of his devise. If others can find better uses for it, his intentions, clearheaded or muddled, are of *mere historical interest (… [so]…)* we can get better grounds for making reliable function attributions (functional attributions that are likely to continue to be valuable interpretation aids in the future) when we ignore (…) 'what [the designer] says' (…) (Dennett: 1990, 186 & 194; quoted earlier in chapter 5).

What Dennett seems to say in this quotation is that original intent does not fix the function of an artifact, and in that respect does not guarantee the *future function* of an artifact. I agree that artifacts can change their function if circumstances change and that they can get better or different functions than originally conceived by the designer. Knowing what the designer had in mind indeed does not fix the future function of the artifact. But it is hard to see why optimality considerations would perform better on this score.

I have considered as a possibility an adaptationist evolutionary macro-approach to artifacts (somewhat along the lines of Preston's theory of artifact selection, see also Lewens: 2002), expecting useless types of artifacts on the long run to die out, and useful types or variants of artifacts to be selected. Not designer intentions, but *selection* determines the function of an artifact because -on the large scale and on the long run- the best designs are systematically picked out. From this perspective, one might say, we see artifacts as adapting to their environment and take on the best possible function. We can thus *predict* that artifacts will evolve to their optimal functions.

The adaptationist approach to technology seems very elegant, especially for Dennett, because it would truly treat technology in the same way as biology, as well as place original intentions *after* selection of the fittest. But I do not believe this is going to help Dennett to his predictive force.

Take the example of the computer that lost its function as a computer, and now functions as an anchor of a boat. Dennett uses this example to claim that intentionalism cannot be right because it fails to predict the future function of the artifact. Now, in order to make the comparison with adaptationism, we would have to rephrase the example, because evolution acts over populations, not individuals. So let's imagine that there was no interest to use computers from the seventies in the nineties anymore as computers, but that some of them did not end up in the garbage, and started to fulfill a new function as an anchor. I can surely imagine a theory of artifacts reasoning somewhat like this, but adaptationism or evolution could not have predicted *that*. It may provide a formal *explanation* of the change in function, but this is still far from a concrete prediction. Adaptation, again, is a contextual notion. An item (or a population of items of a certain type) adapts *to its changing environment*, and thus if Dennett wants his optimality account to be predictive, it would have to predict the direction of this changing environment. And clearly, the optimality account is far from ready to do *that* (just like it is impossible to predict the direction of biological evolution). Neither intentionalism nor adaptationism can predict the future function of an artifact, and probably intentionalism is better able to *explain* changes in function, because it is able to account for changes in the intentional environment. So this attempt seems a dead end.

I have suggested, at the very end of chapter 5, that predictivity is probably to be read in a much more modest, classical and neutral sense of being able to predict the behavior of a functional object in order to test hypotheses about what the object is. This is the sense of 'prediction' that helped William Calvin to his conclusion that the Acheulean hand-axe could never have been a hand-axe: if the item were a hand-axe (hypothesis), it should at least be possible to hold it in your hand (prediction and falsification of the hypothesis). This seems a plausible interpretation of what prediction could mean, but it does not help get a satisfactory answer to the question what Dennett meant when he was writing about optimality being a *more* 'valuable interpretation [aid] in the future'. For why shouldn't and couldn't we want to test hypotheses based on interpretations of the *intended* function of the artifact? Falsification of a theory is not restricted to optimality theories. As a matter of fact, I personally believe that optimality reasoning and reasoning about designer (or) user intent could (and in the hermeneutical sciences *do*) go very well together in a mutually corrective relationship – usually called the hermeneutic circle. (More on that, and on whether Dennett might embrace it, in the conclusion of my dissertation.)

I submit that without further argument from Dennett, the optimality account lacks predictive force with respect to future functions, just as much as an intentional account lacks such predictive force. Alternatively, if we take prediction in its role of falsification of a hypothesis, the optimality account helps generate falsifiable statements just as an intentionalistic account might. Predictive force is thus not a virtue that can *only* be had with the optimality account. Intentionalistic accounts score at least just as good on predictivity. So Dennett at the very least has to be more precise about why his optimality account would have the predictive upper hand.

7.6.2. The practical design stance is intentional

Until now, I have only discussed *scientific* applications of the design stance as method. But as explained in chapter 3, the design stance might also be construed as an *everyday* method of ascribing functions. For Dennett, I have claimed, practical value is an important (perhaps even the most important) value of a conceptualization, and we should therefore consider the possibility that the design stance as construed by Dennett is to be read as an everyday method. Remember the conceptualization of personhood example from chapter 3:

personhood was not to be defined in scientific terms, and thus not to be evaluated as a scientific concept, for the reason that personhood is a practical concept, helping us to make certain moral distinctions. Similarly, the design stance should perhaps be seen as the stance that gives us the practical concept of artifact function.

But the design stance as a practical stance seems very limited as a practical stance. The notion of intention is pervasive in our daily routine with artifacts (cf. chapter 6) and in practical contexts it is also extremely *useful*. This becomes especially clear when we take a look at the possible evaluative role the design stance could have.

The importance of the notion intended function is most clearly illustrated with cases of malfunctioning artifacts. The problem of malfunctioning (analogous to the problem of misrepresentation in the philosophy of mind) is based on the intuition that an artifact might have a function that it is not able to perform. A broken chair is still a chair, and a DVD player with a broken play button is still a DVD player. Intended design accounts of function are traditionally strong to account for malfunctioning, as they frame artifacts in terms of intended function, not actual performance. The optimality account, and the account of the design stance more in general, is per definition unable to account for malfunctioning. It is aimed at what the item at stake is able to do, and not at what it was supposed to do.

'Intention' is intractably connected with attributions of blame and praise (cf. Knobe: 2003, see also p. 35) – as such, the notion of intention will play an important legal or moral role when artifacts fail to fulfill their expected function, or when they work above expectation. Such cases are plentiful in the case of technical artifacts. Think only of badly designed artifacts (may I return it to the store?), artifacts that do not work as expected (should the designer have communicated its function better?), and artifacts that are used in a different way than intended (who is responsible when the sporting-gun is used against humans?), etcetera[86]. These are exactly the contexts in which Dennett's account of the design stance is very weak. Interpretationism has a notorious blind spot for irrationality and suboptimality, and it is to be expected to fail in those cases

[86] Note that I do not wish to imply that the designer of an artifact is always responsible in any of these cases. The claim is more modest that we consider intentions of designers or makers when evaluating such cases.

where we want to distinguish the actual functional capacities of an artifact from the function for which it was designed or used.

7.6.3. The inevitability of intentional reasoning

Intentional reasoning in science will always be looked at critically, but if you *want* to be talking about (technical) functions, you will have to take intentions into account. We do not necessarily need 'strong' intentions but we do need some idea of what the agents whose instruments we are trying to explain needed and wanted, if only because the optimality of a function of an instrument cannot be seen apart from the rationality of its creator or user. Intentional reasoning is part of functional reasoning, or, in Dennettian terms, *but contra Dennett*, the intentional stance is part of the design stance. If this is already so in an intentionality-poor field like archaeology, we may expect intentional reasoning to be even more important in richer fields.

And this is not necessarily unscientific, even though it might never be deemed "real“ science. Science is after the way things are, and whether we like it or not, technical artifacts are created with certain ideas, used with certain ideas, and looked at with certain ideas, which are all very intentional. If we want to study technical artifacts, we will have to consider the question of the role of these intentions. But that *should* not be a problem: if intentions can be described as part of the natural world, just as shoes and cows can, and *if* we can study and talk about intentions, *then* it should at least be possible to study the way people ascribe functions to artifacts when they create, use, and label them. In that respect, I find the naturalistic tendency to deem artifacts improper objects of scientific investigation particularly peculiar: if artifacts are created by minds, and if minds are natural objects, why shouldn't artifacts be proper parts of scientific investigation?

Let me remark that the optimality account *should* be especially good in conceptualizing a methodological notion of function. At least, if we want to evaluate the design stance on the same grounds as the intentional stance. But this is problematic from the outset: what is a (good) scientific notion of function? What would it be good for? To be sure, this is an important question for every account of technical function, and one that is rarely answered. But one would expect a pragmatist with strong naturalistic tendencies like Dennett to be in a good position to answer it.

It was the combination of naturalism and pragmatism in Dennett's philosophy that I found so promising for the Philosophy of Technology. If Dennett were to succeed in constructing a theory of functions with the same force as his theory of mind, the topic might be more easily get the support of other naturalistic-minded philosophers that it is now lacking. And his pragmatism might have helped direct the debates about functions within the Philosophy of Technology more towards the question of what we aim to accomplish with a conceptualization of function – a question that is not really answered, yet, in the Philosophy of Technology. We might then perhaps even dramatically conclude that if even Dennett is unable to give a plausible account of techno-functions, this is bad news for the whole project of conceptualizing technical functions.

Dennett's optimality account is too limited, I think, to cover the whole range of function ascriptions we may want to make, and would be blind for many important *and* practical cases where we want to ascribe functions to artifacts. Whether Dennett likes it or not, often we are especially interested in the (represented or unrepresented) intentions that underlie a certain design. We just may want to know the function the *designer* wanted the artifact to have – irrespective of what happened to the artifact afterwards, and irrespective of some optimal function the artifact could have. In other cases we may be interested in the current intended function – the 'user function'. For instance, when we want to know what the function of an artifact is *now*.

For Dennett, "the intention, if any, with which an item was originally introduced determines, at most, what function the author hoped or intended the item to serve". I agree, but in many cases we may be exactly interested in the function the author hoped or intended the item to serve, and thus will be pressed to find out its underlying intentions.

8 Conclusion: The Interpretation of Artifacts

8.1. Introduction

Our ordinary language and our way of dealing with the world is intentionalistic through-and-through. We see agents as entities that have purposes, ideas, and wishes, and interpret their behavior as such. And in a similar manner we discern objects that have functions, things that are for something and not for something else, *this object is for this, and not for that*. One thing that characterizes us as human beings is that we intentionalize the world.

Intentionalism is a fact of everyday life, and our social and personal survival, as well as many institutions in our society, depend on it. But intentionalistic language is problematic from the start if we try to come to terms with it from a more objective perspective. In everyday life there are many cases when we want to be able to say something determinate about intentions, such as when we want to decide whether we should speak of murder or a fatal accident, of a lie or a mistake. This is notoriously hard because we have to work with indirect and circumstantial evidence. And in science, intentional language is still, at best, treated with a lot of caution – and for good reason.

I have presented Dennett's intentional systems theory very sympathetically as a minimalistic account that does the required justice to intentional language, in such a way that it can even be sufficiently acceptable to science, without being naïve about the limitations we face. Dennett makes a number of pragmatic choices that amount to, I have claimed, a refreshing, comprehensive and unificatory account of intentional states: Intentional Systems Theory. Intentional Systems Theory is liberal about our ascribing intentional states to all sorts of agents; a liberalism that is backed up by a promising biological take on issues of the mind.

Dennett's development of the design stance as part of the theory of the stances is, in the same vein, liberal and refreshing as far as it concerns biology. The idea that we can see biological evolution as a design process, without a designer, is original and provocative. But when it comes to the *technical* functions, Dennett suddenly backs out. He gets very cautious and strict about intentional notions of functions, and it often seems that Dennett is less hesitant to grant intentional ascription rights to 'Mother Nature' than to human designers of a technical item. In the attempt to treat biology and culture on a par, it seems that culture, being last in line, had to budge in order to save the larger bionaturalistic picture. It is then almost paradoxically cruel that the resulting picture of technology has to be rejected all the same on the basis of internal, naturalistic standards. Does the Dennettian paradigm result in contradiction on either side of the fork, or is there room for reparation? To answer this question, let me go through the whole argument of the dissertation, but now up-tempo.

8.2. The argument against the optimality account

8.2.1. The question

Conceptualization of artifacts and their functions is a much ignored but intuitively important philosophical topic. Especially naturalists, by nature suspicious when it comes to intentionality and mind-dependent phenomena, show little enthusiasm to pick up the topic. Dennett is an exception, and as an important representant of the naturalistic tradition, and deviser of an original theory about the validation of the use of intentionalistic terminology, he is the candidate *per excellence* to shed naturalistic light on the question whether artifact functions are a proper part of a (naturalistic) philosophical theory.

Does Dennett's method of the stances provide an adequate framework for conceptualizing functions of artifacts?

This was the question I set myself. It is a specification of the larger issue of what we mean when we talk about artifact functions and what the conditions are under which we can legitimately ascribe functions to artifacts. The question can be answered in a lot of ways, if only because philosophers disagree about what concepts are, what a proper conceptualization is, whether and why it is important, and how conceptualization relates to attribution. Some philosophers would say that conceptualization and attribution are really three different issues (by and large, an ontological, a conceptual and an epistemological one), others, such as Dennett, believe that they all boil down to the same.

The question what an "adequate" framework for conceptualizing and ascribing functions to artifacts is, cannot be seen apart from the larger question why we *should* be looking for such an adequate framework. Who cares when we make a 'mistake'? Who's to judge when our application of a concept is 'correct', and on the basis of what standards? These are all huge and important philosophical questions. I have chosen to limit myself to Dennett's pragmatic approach to such questions (3.5). Thus, the specified question of thesis becomes a more narrow:

Does Dennett's method of the stances, according to his own standards and criteria*, provide an adequate framework for ascribing functions of artifacts?*

I have explained in chapter 3 (see also 2.2) that a good or useful concept of a term, according to Dennett, equals the concept as it would be if it were part of a good theory of *attribution*. A good theory, then, for Dennett, is a practical theory, that is to say, it is objectively verifiable, applicable and useful. It may help us effectively explain things, or to predict future events. It is, for Dennett, also a third-person perspectivist theory, exactly because first person perspective reports are unreliable and very hard to work with and thus fail the pragmatic criterion (see chapter 3).

The pragmatist says that a good concept is a practical concept, something you can work with, something that helps you solve problems. Settle issues of responsibility, for instance, or help categorize items. This certainly is not necessarily unscientific: correspondence of your concepts to the world is very useful (e.g. it gives better predictions) and will therefore usually be preferred by the pragmatist, but correspondence is not a goal *an sich*. For Dennett, science is certainly a very practical way to define and investigate concepts and as such the main candidate to do it. We may even say that Dennett takes some distance from philosophy, towards science, leaving slightly open what the proper role of philosophy should be.

Pragmatism in the form Dennett defends it is thus compatible with science, and Dennett can indeed be seen as a naturalistic pragmatist. The pragmatist is not committed to finding answers through science. In some cases, there may be other, better (more practical/useful etc.), ways to find answers. 'Adequate attribution' is for the pragmatist relative to a specific purpose or a certain practical need we have to answer certain questions. Sometimes, science can help us to answer such questions, in other cases we may need to find another way to deal with them. I have discussed an example of an un-scientific kind of conceptualization in chapter 3 (the notion of personhood). Nevertheless, Dennett is most interested in scientific applications of terms, and clearly his naturalistic tendencies dominate his views. We may expect Dennett to come up with a naturalistic notion of artifact function, that as such has practical worth.

8.2.2. The two standards

Dennett's ideas on technical artifacts must be seen in the light of the larger theory of the stances, and the theory of the stances, in turn, is part of even a larger philosophical project that I have characterized as pragmatic naturalism.

Dennett wants a 'grand theory' of the mind, a theory that spans the whole mind, from content to consciousness, from the first signs of mental life in animals to the complex operations of human minds, with extensions to issues that are related to issues of mind, such as religion, ethics, freedom and of course technology. He wants to connect these issues such that they make up a continuous view. I have derived two central standards from this view: *empirical correctness* and *methodological utility*. To be sure, both have plausibility of their own, that is, there is good reason for many to accept them as standards, although people may disagree about their weight. But they are really central and crucial to a theory like Dennett's because these standards characterize his philosophical project all the way through.

By *empirical correctness* I mean that the conceptualization, or the rules of attribution, of a certain concept must do justice to the relevant empirical facts. For Dennett, I have argued, this means that the theory should roughly describe how we ascribe intentions and functions to entities. This is clearly a naturalistic requirement. The naturalist, or many at least, prefers to go as little beyond empirical reality as possible. If you want to know what something is, the naturalist says, you study it systematically in the empirically describable world, or you phrase it as a theoretical concept that helps us organize certain knowledge. Many naturalists prefer to stick to a description of a certain phenomenon (e.g. what people mean when they use a certain term), and leave normative issues at rest.

But a mere description of intentional facts is not enough for most philosophers. What if we make systematic mistakes in ascribing attitudes and functions? And what counts as 'wrong' in this case? On the basis of what facts can we determine that our ordinary ascriptions are right or wrong? These kinds of *normative* questions interest the philosopher, and Dennett wants to answer them as well. In the end, IST is a methodological (and thus normative) theory of how we *should* apply intentional terms in science and a justification of such applications. Dennett believes that the descriptive version of the intentional stance needs some revision. This revision, to be sure, concerns especially the *ontological status* of these attitude terms (3.6). He, then, chooses a pragmatic way of distinguishing good ascriptions from bad ascriptions. Good ascriptions are (methodologically) practical ascriptions. I call this the *methodological standard*.

Dennett's normative ambitions create a certain tension in his work between his desire to stick close to the empirical facts, and his methodological, revisionist, ambitions. It is a widely shared view that norms cannot be simply derived from facts. Normative statements can therefore not be founded in is's, or, as the naturalist prefers, they should not be made at all. So, on the one hand we need a normative theory that is outside empirical reality, to evaluate the empirical facts with; on the other hand, the normative theory cannot stand apart too much from the rest, because a free floating normative theory is unwelcome in a any naturalistic project.

It is therefore important that Dennett connects in some way the empirical criterion with the methodological theory, for instance by showing that the empirical facts in some sense support the methodological claim. And this is exactly what Dennett seems to want to do in his work on the intentional stance. The "empirical" intentional stance is a real-life fact, almost hard-wired in our brains. It has evolved over a long period of time, which explains that and why it works so well. This, in its turn, provides at least prima facie evidence that is also a good method. Put differently: that we use the intentional stance so much, and that we have done that in our whole evolutionary history (see Dennett: 2006), is most plausibly explained by the fact that it works (why and how is another issue). And a stance that works has good references when turned into a scientific method. Thus, careful normative conclusions (the method) are drawn on the basis of the pragmatic value of taking the intentional stance (the actual facts).

With respect to the evaluation of the design stance, then, Dennett has to show that his account prescribes a *good* way to ascribe functions to artifacts, while making sure that these recommendations do justice to and are in some way intelligibly connected to the actual empirical facts, i.e. the way we actually ascribe functions to technical artifacts. Applied to the interpretation of artifacts, then, a good conceptualization (i.e. attribution) of functions of artifacts:

(A1) gives a good description of the way we ascribe functions to artifacts.

(A2) gives a good method for ascribing functions to artifacts, scientific or otherwise.

8.2.3. The optimality account

The key question is: how *does* Dennett propose to conceptualize artifacts and their functions? The answer is not so easy, because Dennett is more eager to write about biological functions, than he is about technical functions. His ideas on the interpretation of technical artifacts *specifically* are phrased in one single paper; the rest of it has to be carefully derived from his theory of the stances.

The attempt to create a non-intentional account of function can be explained by Dennett's attempt to bring biology and culture closer in line with each other. Biological function explanations have to look like technological function explanations, and vice versa (principle of function generality). This move to treat biofunctions like technical functions is attractive, original, and, liberating – and fits qua style Dennett's philosophy very well. Function generality can, however, be read in two ways: to treat functions in an equally intentional way, or to treat them in an equally non-intentional way. In his early work on the intentional stance, Dennett connects design interpretation to designer *intent*, especially in the biological realm, suggesting that he wants a generic account of *intentional* functions. But Dennett is *against* an intentionalistic account of technical functions. Furthermore, Dennett's naturalistic principle of design primacy, that says that intentions are explained in terms of design, forbids an intentional account of design - in order to prevent circularity between the design stance and the intentional stance. This pushes Dennett in the direction of a *non-intentionalistic* optimality account of functions, defined as follows:

The function of a certain item is – or should be – what it is best able to do (or be) given its physical constitution and its context, and not what the designer(s) or user(s) intend it to do (or be). Function interpretation, thus, is not –or should not be- a matter of interpretation of intentions.

The optimality account is *explicitly* put in contrast with intentionalistic approaches, both function accounts that refer to original designer intentions *and* function accounts that refer to user intentions. This means that designer intent can be overruled by optimality considerations, and that intent is the conclusion of an inference about a design, and not its starting point.

The optimality account basically says that the function of an artifact is *what it is best able to do* ("it counts against something being a cherry pitter if it would be a demonstrably inferior cherry pitter") in a given context. Optimality, personified perhaps as a hypothetical rational designer, then figures as an assumption that guides our reasoning about what an artifact('s function) is. For as far as this designer plays a role, it is that we might infer what 'the designer must have meant' or 'what was apparently intended' on the basis of optimality considerations. The claim is that the optimality account is predictively stronger, and that it better captures functions of artifacts.

8.2.4. Dennett's arguments against intentionalism

Being influenced by the argument of the Intentional Fallacy, Dennett raises two problems of intentionalism. The rejection of intentionalism is Dennett's prime argument for the optimality account.

The first problem of intentionalism, according to Dennett, is that specifications of intentions, including those of artifact designers of course, are indeterminate and unreliable (indeterminacy argument). Dennett concludes that the intentionality account fails and that we should therefore choose his alternative. The argument is self-undermining. That interpretations of intentions can be indeterminate and unreliable is not a reason to reject an intentionalistic account, at least, cannot be a reason for Dennett to reject it. That would undermine his own intentional systems theory, that has (unreliable and indeterminate) intentional ascriptions at its *core*.

The second problem of intentionalism is that designer intentions are irrelevant for our function attributions. For instance, artifacts may acquire a new function, not intended by the designer. From this, Dennett concludes again that the intentionality account fails and that we should choose the optimality account. Dennett's second argument fails as well, this time because it barks against the wrong tree. It only works against forms of extreme intentionalism that take designer intentions to be the sole determinant of the function of an artifact. This is indeed an implausible form of intentionalism that is hardly defended anymore. It is implausible, first, because intentions are given supernatural powers (my intending the tea pot to be an airplane does not make the tea pot an airplane) and because artifacts can and do acquire functions that the designer never intended. The argument does however not yet reject all forms of intentionalism. There are more versions of intentionalism to be discussed, amongst which modest actual intentionalism and hypothetical intentionalism (see p.118), so Dennett cannot present his optimality account as the next best thing.

Dennett presents his optimality account as the best and only alternative for the failing intentional account. But because he has not successfully rejected all forms of intentionalism, his optimality account cannot be presented as the best alternative. He will either have to find better reasons why we should reject intentionalism or, preferably, come up with sufficient *positive* reasons to choose the optimality account.

What could such positive reasons be? We find in his work two arguments for the optimality account: it may give a better description of reality, and it may yield better predictions (or more generally, have greater methodological value). Both arguments can be evaluated by appeal to the two standards that I defined for evaluating Dennett's theory.

8.2.5. The empirical evaluation

When discussing the optimality account, Dennett often suggests that his optimality account is descriptively most accurate, usually by reference to reverse engineering or scientific method (archaeology). But this is not conclusive: even *if* these cases describe function attributions on the basis of optimality considerations alone, which I doubt, why should these be the relevant cases to describe? There may certainly be cases where the optimality account gives the best description (for instance when we want to find out for what purpose we may best use a certain artifact), but there are plenty of cases where the optimality account does not give the better description. For instance, when an object is designed for a certain purpose, but is not able to fulfill that function due to accidental circumstances, subjects still categorize that object in terms of its intended function, rather than in terms of an alternative function that the object would currently be able to perform well.

In fact, empirical research so far points out that we usually do *not* conceptualize artifact functions in terms of optimality. At least in everyday life, our concept of artifacts, and our categorization of them, seems to be primarily framed around original designer intentions. Interpreting functions of technical artifacts derives from our more general ability to interpret intentions of people, i.e. when we interpret an artifact, we usually interpret the intentions of its designer. This is in accordance with the general idea of the design stance: if you know the purpose, you know what it will do. But 'purpose', in these studies, turns out to be intended purpose, and not "optimal purpose". It seems that the step towards a non-intentional design stance, was not so fortunate.

8.2.6. The methodological evaluation

Dennett writes that "we can get better grounds for making reliable function attributions (functional attributions that are likely to continue to be valuable interpretation aids in the future) when we ignore (...) 'what [the designer] says'" (Dennett: 1990, 194). I take this as a clear hint that Dennett's optimality account is meant to perform better as a method to attribute functions to an artifact.

Dennett's gradual and generous account of *attitudes* is well supported as a method of ascription by the cognitive ethological literature, much better than many alternative philosophical conceptualizations of intentional terms (e.g. the armchair distinction between behavior and action). This gives further support to the intentional stance as methodology for ascribing attitudes, as well as to the theory of mind that follows from it. The design stance, underdeveloped as it already is, would profit a lot from such methodological support. But even in a science like archaeology where it should do very well, it is not convincing.

There are not many sciences about artifacts and their functions, but archaeology is a welcome exception to test Dennett's optimality account. And it should be convenient for Dennett, because, at first sight, the optimality account has much to say for itself in this domain. By scarcity of knowledge about intentions of our ancestors, the optimality account seems a reasonable alternative to say something after all about artifact functions. But it turns out that even in archaeology, some questions are just better resolved by (additional) appeal to ancient intentions, even if our knowledge of it is scarce, speculative and indirect. The reason for this lies in the fact that archeologists, for as far as they don't want to be behaviorists, are not after just any good use for a particular object, but after original intended functions of artifacts. The optimality account may well tell us how to put an ancient object to good use, but this can amount to a very long list of good uses (an old hand-axe can be a museum piece, a paper weight, a piece of decoration, a weapon), and taking into consideration the intentional context in which it was created may significantly narrow down the options about the object's history.

Also when Dennett's account is construed mainly as a predictive theory, it remains unclear what the optimality account would predict. The optimality account is just as unable to say something about the future function of an artifact as an intentionalistic account, if we would even care for one. The optimality account might generate good hypotheses, but intentional accounts can do that just as well.

Methodological value can also lie in 'everyday' methodological value. Dennett might show that a conceptualization of function without reference to designer intent helps, for instance, solve legal problems. But, as one might expect after the results on the empirical score, the everyday notion of technical artifact function is intentionalistic through-and-through.

8.3. The conclusion

I conclude that the optimality account has limited applications and would certainly improve if it were able to take into account information about the intentional context in which it was created or used and its intended design. Dennett's optimality account of the design stance has a lot of potential, but it fails to prove itself .

The theory of the intentional stance offers, I think, an inspiring way of conceptualizing attitude terms such as belief, intention and desire. The intentional stance at least has, in my view, shown its worth. It is minimalistic enough to be realistic about what we can say about the ontology of attitudes, yet tries to maximize its potential by sticking close to the empirical facts and emphasizing the methodological virtues of the stances. The result is a liberal account of attitudes, developed with a keen eye towards practical use.

The idea of a design stance as a certain way of looking at certain objects that has both concrete empirical roots as well as methodological potential, is by itself a promising tool to get more grip on the notion of artifact function. As such, it is an interesting concept for the Philosophy of Technology, especially because it is embedded in a larger theory of mind and thus offers a close connection to a well established field in philosophy. Its pragmatic underpinning may make the design stance more directly relevant and interesting for an audience of engineers.

The strength of the intentional stance is that it provides a way to say something about actions from the perspective of the agent, yet from the third person perspective. The design stance, being set up to bridge the gap between mind and world, a gap that is particularly important for the philosophical theory of technical artifacts, could similarly say something about the thoughts behind an artifact, and thus go beyond a mere physical description of mechanisms, without being committed to strong or subjective claims about (determining) intentions.

Promising as it is, the way Dennett develops his account of the design stance, in terms of optimality-not-intentionality, just fails to be convincing. Dennett's negative arguments against intentionalism fail to hit the relevant target, and rather backfire on Dennett's own theory of intentional states.

In addition, the optimality account is not an accurate description of function ascription to artifacts. Human beings see intentions everywhere, certainly in technical artifacts. The "intentional fallacy" simply is a plain fact of human life. The optimality account just does not provide a clear method and acclaimed advantages (e.g. prediction) can be had by intentionalistic accounts as well. Intentionalistic accounts are also stronger if we are looking for conceptualizations that could be of use in everyday life (e.g. in law or ethics) – such conceptualizations most clearly need to be connected to our everyday way of conceptualizing artifacts. But it also fails to convince as a scientific method, which is slightly paradoxical because to a large extent the optimality account fails because Dennett is too scientistic about it (i.e. tries to eliminate references to intentions in the technical design stance).

Dennett wants a theory of function that has as little reference to intentional states as possible. I agree with Dennett that the philosopher of technology should be careful not to misconstruct the history of technology as a history of creative geniuses (cf. Lewens: 2004). The history of artifacts is not a history of explicit, represented intentions. Many, if not most, artifacts that we know and use every day, are not the result of an explicit, creative intention of a designer, and not every designer intention leads to a successful artifact. Not only users, but also economy, culture, politics, to name a few factors, play a big role in the creation of artifacts, and co-determine what the artifact is. Marketing is the greatest artifact-baptizing industry there is!

Artifact creation is usually the result of a collective process (Gibson and Ingold: 1994). It may already be hard to determine the intention of an individual – but even harder to determine the intention of a *group* of individuals, and virtually impossible if they are relatively unorganized and scattered over time and space. An individual designer may create an artifact without having a clear purpose in mind (the stopper of a beer bottle was, according to Dutch advertising at least, invented mindlessly by a guy that dreamed of a future with his beautiful girlfriend, while working in a factory). We may be even more doubtful as to whether the collective intentions of a group of designers are represented somewhere. This is more a problem for the internal realist and for the first person perspectivist, who will find it difficult to accommodate collective attitudes in their theory of mind. Dennett's third person perspective interpretationism has relatively little problems of ascribing attitudes to groups, because it is very liberal with respect to the entities to which we can ascribe attitudes (see Tollefsen: 2002 for an argument that the intentional stance can indeed be successfully applied to organizations).

What I find elegant of Dennett's theory of the stances, is that (successful) ascriptions have (quasi)ontological results. From that perspective, I understand Dennett's choice *for* the optimality account. The proof of the pudding is in the eating, so if we want to baptize an artifact as an x (or if we want to claim that particular object has function x), it better behave as an x.

So I fully agree with Dennett that it would be naïve to claim that artifacts can be understood by designer intention alone. But this should not lead us to the other extreme. An account of artifact function that has no reference to intentions whatsoever is equally implausible.

Intentional accounts of artifact functions allow us to say something about the ideas behind an artifact – which is, I think, an important fact about an artifact if we want to explain its coming into existence. Another advantage is that a more intentionalistic account has more opportunity to deal with cases in which our ideas about an artifact do not match the world, for instance, when an artifact has capacities we are unaware of or when it does not work as we want it to.

Finally, an account of the design stance that takes into account intentions matches our ordinary way of looking at technical design. I should note here that I believe that conceptualizations that aim to stick close to ordinary language use, should try to work less from armchairs, and work closely together with empirical researchers. I consider this one of my major conclusions in chapter 6 and am confident that there is a world to be won in that area.

The optimality account is too rigid. Dennett's liberal take on intentions provides a considerable respite, it is a pity that Dennett must be so strict about intentions in function ascriptions. As such, the optimality account just does not fit the larger theory, which should worry Dennett a lot, because he wants a unified theory of mind and world that *should* include a neatly fitting account of culture and technology. Ironically, naturalistic considerations pushed Dennett towards non-intentionalism. And now, naturalistic considerations - in the form of the two standards - push it right back.

Another irony is that with his intentional stance Dennett has all the instruments he needs to do justice to intentional states without being committed to heavy duty claims about what they are, but is no at liberty to use it. The interesting question now is: Would that be Dennett's favorite approach to artifact functions, were it not for the circularity that 'forced' him to the optimality account? Recall that the optimality account was a solution to the problem of circularity between design and intent (5.6), but leads to internal problems elsewhere in the theory *and* an implausible account of artifact functions. Obviously a bad result for Dennett, but if we look at it from the positive side, Dennett can now choose his poison: intentionalizing his account of the design stance, on pain of circularity; or sticking to optimality, on pain of internal instability and implausibility… From this perspective, intentionalizing may be the better choice after all.

8.4. Discussion: artifact interpretation revisited

Dennett naturalized his design stance, but let's now look at the prospects of (re)intentionalizing it. Were it not for the circularity between the design stance and the intentional stance, it seems that Dennett should be perfectly able to account for a notion of artifact function that could do justice to intentional facts about artifacts; his own intentional stance methodology not only paves the road for it, it is even jeopardized when Dennett would stick to his criticism on intentional approaches to artifact function. Dennett's own argumentation against intentionalism is strictly directed towards strong intentionalism, so a mild intentionalist account of function may fit Dennett's framework just fine. Even better, his paradigm *should* be especially able to capture the relevant intentional facts behind artifacts.

One might even wonder, given the results of this dissertation, whether the optimality account *has* been the proper interpretation of Dennett after all, given such contradictory results. After all, the decision to interpret Dennett's design stance as the optimality account, was at least partly based on the fact that it would otherwise lead to contradictions within the larger theory of the stances. But the optimality account, that was supposed to solve a threatening contradiction, turns out to be incompatible with the larger theory (although on a more general level) as well - doesn't the principle of charity demand that I revise the interpretation? This worry goes right into the heart of the problem of interpretation, but I believe it does not undermine the general conclusion of this thesis.

Suppose, for the sake of the argument, that I misunderstood Dennett. The message of this dissertation would then still be that Dennett's theory of the design stance as it stands leads to conflicts with the larger theory under both the optimality interpretation (conflict with the two standards) and the intentionalistic interpretation (conflict with the principle of design primacy). The principle of charity may demand that a theory must be interpreted such that the result is coherent and consistent, but obviously there are limitations to this requirement (ordering whole papers to be removed from the view is clearly beyond the rights of the interpreter, I would say). Interpretations have to aim for consistency, but authors are not always as consistent in their thoughts. The result of an interpretation may be that given the available sources, no stable interpretation is possible. It is just a inherent characteristic of the project of interpretation that one might have to wonder whether the interpreter has made a mistake when an interpretation does not come off the ground as a stable whole, or that it was the author who made the mistake.

If I am correct that Dennett's design stance, on either side of the fork, leads to conflicts in the larger theory, what side of the fork would he prefer? I personally prefer an intentionalistic account of functions, but am at the same time quite convinced that Dennett would choose to stick to the optimality account if he had the choice. Would a philosopher who has made fame arguing that thermostats can think, that religion is like a virus infecting the heads of people, and that nature is like an intentional agent, find any challenge in claiming that artifact functions are intentional? I believe that Dennett himself would find an intentionalistic account of technical functions just too boring. We might say, as a variant on Dennett's cherry pitter claim that it counts against something being a good interpretation of Dennett if it would be a demonstrably *boring* interpretation of Dennett. This somewhat psychological and unconventional standard of interpretation can obviously not carry any substantive weight in a philosophical interpretation, but I did want to just mention it as an extra consideration worth letting pass the readers mind in this discussion.

Contrary to the optimality account that is *defined* as an anti-intentionalistic account, an intentionalistic account could combine optimality considerations with intentional information. I personally would favor an account where both are combined, reflexively, and enhancing each other (the hermeneutic circle). Such a combination would cover a much wider range of cases in which we might want to know more about a function of an artifact, especially when we want to know more about its actual functional history. Whenever we have reason to believe that there is more to the artifact than merely its optimal design (think of the Babylonian battery), we could go deeper and embed our theories of what the artifact is or was supposed to be in a richer intentional context and ask the question: "Could this have been what they had intended"? Conversely, hypotheses about intended design can be tested with optimality considerations: "This can never work, they must have had a different purpose". We could think of a positions such as modest actual intentionalism (the purpose of an artifact is what the designer intended it to be, unless it can be shown that this purpose is not attainable by the artifact), or hypothethical intentionalism (the purpose of an artifact is what the designer most plausibly must have intended with it), as shortly discussed in 5.8.

It has not been the project of this investigation to provide a precise theory of what a more intentionalistic account of the technological design stance would look like, but I would like to spend the last pages of this dissertation by shortly sketching where I believe reparations should be made if such an account were to be developed, under the constraint that they should be acceptable within a Dennettian framework.

I see two big problems that need to be accounted for in order to steer the optimality account into a more intentionalistic direction; both follow logically from chapter 5:

(1) An intentionalistic technical design stance goes against the principle of function generality (IA3), because technical designs are then treated different than biological designs;

(2) An intentionalistic technical design stance goes against the principle of design primacy (IA2), because the design stance is supposed to explain intentions, not be explained by intentions.

The problem lies, I think, mainly in (2). The problem of failing function generality does not come up if we construct the technical design stance as proposed above. Recall that Dennett's original account of the design stance was very generous in ascribing intentions to nature. His adaptationism is, or was, quite liberal in its permission to ascribe intentions to nature's hypothetical designer, Mother Nature. His notion of biological function permitted taking the intentional stance "towards Mother Nature" (hence: reasoning about intentions). Obviously, in order to align biology and technology, the notion of technical function should be devoid of quests for "real" intentionality. But, as I have argued before, Dennett could use his own views on intentionality to account for the intentional nature of technical artifacts. So I see little problems in re-importing the intentional stance for biodesigns *per se*.

The real problem, I submit, lies in (2). It was the principle of design primacy that caused a circularity, and pressed Dennett to optimality. The problem is that the principle of biological naturalism says that when our intentional explanations fail, we should take the design stance. So, in the end (when the intentional stance fails), intentional events have to be understood biologically, or as products of biological evolution. This means, for Dennett, that the intentional stance is, eventually, reduced to the (biological) design stance. If Dennett wants to hold on to his claim of biological naturalism, the design stance cannot in its turn refer to the intentional stance, on risk of circularity.

Let me remark that on second thought, the thesis of function generality is not well met under the optimality account. Dennett wants a design stance that is equally applicable to biology and nature, and in its general formulation it is, but when push comes to shove, it is the biological design stance that really counts; counts in the sense of *explains*. Dennett's construal of the design stance in terms of optimality is clearly driven by the ambition to save the *biological* design stance, so that it can do the explanatory work it is supposed to.

My (hopefully not surprising) diagnose thus is that the pain lies in the problem of circularity and the ambition to reduce intentionality to design. Might it be solved? Perhaps. Dennett might try to turn the apparent circularity in a virtuous progress (as opposed of a vicious regress). Interpretation is, perhaps, essentially circular, 'hermeneutically circular' so to speak, so perhaps Dennett should just accept that the design stance and the intentional stance rely so much on each other, as facts that are just part of their being interpretative stances. Isn't this exactly the point of Quine's example of a lone traveler that tried to translate the utterances of the inhabitants of another land? Starting from scratch, and no certain starting point at all, he had to interpret his way into the native's language. Interpreting designs and minds is perhaps not so much different. Starting with rough hypotheses about intended purpose, capacity, and of course, assumptions of optimality and rationality, we slowly might get to a better understanding of what an artifact could have been made for, and what other functions it may afford. I believe that connecting the design stance and the intentional stance might in fact do more justice to their (sometimes confusing) relatedness. Whether this works, and indeed effectively ends the circularity, I'm not sure, but it might be worth a shot.

Putting the intentional stance back in line with the design stance would put the intentional stance at the core of the theory of the stances. This may conflict with bio-naturalistic ambitions, but the current account does not sit easy with bio-naturalistic ambitions either. Intentionalizing the technical design stance thus might not make matters worse for Dennett; on the contrary, it might solve a number of important problems.

Obviously, the result will never be a die-hard scientific account of functions, but if Dennett had wanted to opt for that, he would not have bothered to construct the theory of the stances anyway.

References

Allen, C. and Bekoff, M. (1997) *Species of Mind; The Philosophy and Biology of Cognitive Ethology,* MIT Press, Cambridge MA.

Anscombe, G. E. M. (1957/1963) *Intention,* Harvard University Press, Cambridge/London.

Audi, R. (ed.) (2001) *The Cambridge Dictionary of Philosophy,* Cambridge University Press, Cambridge.

Baker, L. R. (1987) *Saving Belief; A Critique of Physicialism,* Princeton University Press, New Jersey.

----- (1995) *Explaining Attitudes; A Practical Approach to the Mind,* Cambridge University Press, Cambridge.

----- (1998) 'The First Person Perspective: A Test for Naturalism', in: *American Philosophical Quarterly* 35: 327-348.

----- (2000) *Persons and Bodies; A Constitution View,* Cambridge University Press, Cambridge.

Baker, L. R. (forthcoming) *The Metaphysics of Everyday Things.*

Baldwin, D. A. (1992) 'Clarifying the Role of Shape in Children's Taxonomic Assumption', in: *Journal of Experimental Child Psychology* 54: 392-416.

Beardsley, M. (1970) *The Possibility of Criticism,* Wayne State University Press, Detroit.

Bechtel, W. and Richardson, R. C. (1992), 'Emergent Phenomena and Complex Systems', in: Beckermann, A., Flohr, H., and Kim, J. (eds.), *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism,* Walter de Gruyer, Berlin: 257-288.

Bloom, P. (1996) 'Intention, History, and Artifact Concepts', in: *Cognition* 60: 1-29.

Bratman, M. E. (1987/1999) *Intention, Plans, and Practical Reason,* CSLI Publications, Stanford, CA.

Carroll, N. (1999), 'Interpretation and Intention: The Debate between Hypothetical and Actual Intentionalism', in: Margolis, J. and Rockmore, T. (eds.), *The Philosophy of Interpretation,* Blackwell, Oxford: 75-95.

Casler, K. and Keleman, D. (2005) 'Young Children's Rapid Learning about Artifacts', in: *Developmental Science* 8 (6): 472-480.

Chaigneau, S. E. (2002) *Studies in the Conceptual Structure of Object Function,* Ph.D. Thesis. Emory University Atlanta.

Cherniak, C. (1986/1992) *Minimal Rationality,* MIT Press, Cambridge/London.

Child, W. (1994) *Causality, Interpretation, and the Mind,* Oxford University Press, Oxford.

Csibra, G., Bíró, S., Koós, O., and Gergely, G. (2003) 'One-year-old Infants Use Teleological Representations of Actions Productively', in: *Cognitive Science* 27: 111-133.

Csibra, G. and Gergely, G. (1998) 'The Teleological Origins of Mentalistic Action Explanations: A Developmental Hypothesis', in: *Developmental Science* 1: 255-259.

Csibra, G., Gergely, G., Biro, S., Koos, O., and Brockbank, M. (1999) 'Goal attribution Without Agency Cues: The Perception of "Pure Reason" in Infancy', in: *Cognition* 72 (3): 237-267.

Cummins, R. (1975) 'Functional Analysis', in: *Journal of Philosophy* 72: 741-765.

Davidson, D.----- (1980/2001) *Essays on Actions and Events,* Oxford University Press, Oxford.

----- (1982/2001), 'Rational Animals', in: *Subjective, Intersubjective, Objective,* OUP, Oxford: 95-105.

----- (1984/2001), 'Radical Interpretation', in: *Inquiries into Truth and Interpretation*, Oxford University Press, Oxford: 125-139.

----- (1994/2001), 'Thought and Talk', in: *Inquiries into Truth and Interpretation,* Oxford University Press, Oxford: 155-170.

Davidson, I. and Noble, W. (1994), 'Tools and Language in Human Evolution', in: Gibson, K. and Ingold, T. (eds.), *Tools, Language and Cognition in Human Evolution*, Cambridge University Press, Cambridge: 363-388.

Dawkins, R. (1976/1989) *The Selfish Gene,* Oxford University Press, Oxford.

Defeyter, M. A. and German, T. (2003) 'Acquiring an Understanding of Design: Evidence from Children's Insight Problem Solving', in: *Cognition* 89: 133-155.

Dennett, D. C. (1969/1993) *Content and Consciousness,* Routledge, London / New York.

----- (1978a/1998a) *Brainstorms; Philosophical Essays on Mind and Psychology,* MIT Press, Cambridge/London.

----- (1978b/1998b), 'Conditions of Personhood', in: *Brainstorms; Philosophical Essays on Mind and Psychology,* MIT Press, Cambridge/London: 267-285.

----- (1978c/1998c), 'Intentional Systems', in: *Brainstorms,* Bradford Books, Cambridge/London: 3-22.

----- (1978d/1998d), 'Skinner Skinned', in: *Brainstorms; Philosophical Essays on Mind and Psychology,* MIT Press, Cambridge/London: 53-70.

----- (1983) 'Intentional Systems in Cognitive Ethology; The 'Panglossian Paradigm' Defended', in: *Behavioral and Brain Sciences* 6: 343-390.

----- (1987a/2002a), 'Evolution, Error, and Intentionality', in: *The Intentional Stance,* MIT Press, Cambridge/London: 287-321.

----- (1987b/2002b), 'Intentional Systems in Cognitive Ethology; the "Panglossian Paradigm" defended', in: *The Intentional Stance,* MIT Press, Cambridge/London: 237-286.

----- (1987c/2002c), 'Reflections: Real Patterns, Deeper Facts, and Empty Questions', in: *The Intentional Stance,* MIT Press, Cambridge/London: 37-42.

----- (1987d/2002d) *The Intentional Stance,* MIT Press, Cambridge/London.

----- (1987e/2002e), 'Three Kinds of Intentional Psychology', in: *The Intentional Stance,* MIT Press, Cambridge/London: 43-68.

----- (1987f/2002f), 'True Believers', in: *The Intentional Stance,* MIT Press, Cambridge/London: 13-35.

----- (1987g/2002g), 'Reflections: Interpreting Monkeys, Theorists, and Genes', in: *The Intentional Stance,* MIT Press, Cambridge/London: 269-286.

----- (1990) 'The Interpretation of Texts, People and Other Artefacts', in: *Philosophy and Phenomenological Research* 1 (Supplement): 177-194.

----- (1991a) *Consciousness Explained,* Back Bay Books, Boston (etc.).

----- (1991b), 'Two contrasts: folk craft versus folk science, and belief versus opinion', in: Greenwood, J. D. (ed.), *The Future of Folk Psychology,* Cambridge University Press, Cambridge: 135-148.

----- (1993), 'Quining Qualia', in: Goldman, A. (ed.), *Readings in Philosophy and Cognitive Science*, MIT Press, Cambridge, Massachusetts: 381-414.

----- (1995a), 'Back from the Drawing Board', in: Dahlblom, B. (ed.), *Dennett and his Critics; Demysifying Mind*, Blackwell, Oxford/Cambridge: 203-235.

----- (1995b) *Darwin's Dangerous Idea; Evolution and the Meanings of Life,* Touchstone, New York.

----- (1996) *Kinds of Minds; Towards an Understanding of Consciousness,* Basic Books, New York.

----- (1998), 'Self-Portrait', in: Dennett, D. C. (ed.), *Brainchildren; essays on designing minds,* MIT Press, Cambridge: 355-366.

----- (1999), 'Real Patterns', in: Lycan, W. G. (ed.), *Mind and Cognition; An Anthology*, Blackwell Publishers, Oxford: 100-114.

----- (2000), 'With a Little Help from My Friends', in: Ross, D., Brook, A., and Thompson, D. (eds.), *Dennett's Philosophy; A Comprehensive Assessment*, MIT Press, Cambridge/London: 327-388.

----- (2003) *Freedom Evolves,* Viking Press, New York.

----- (2005) *Sweet Dreams; Philosophical Obstacles to a Science of Consciousness,* MIT Press, Cambridge/MA.

----- (2006) *Breaking the Spell; Religion as a Natural Phenomenon,* Viking, New York.

Diesendruck, G., Markson, L., and Bloom, P. (2003) 'Children's Reliance on Creator's Intent in Extending Names for Artifacts', in: *Psychological Science* 14 (2): 164-168.

Dipert, R. R. (1993) *Artifacts, Art Work, and Agency,* Temple University Press, Philadelphia.

DiYanni, C. and Keleman, D. (2005) 'Time to Get a New Mountain? The Role of Function in Children's Conceptions of Natural Kinds', in: *Cognition* 97: 327-335.

Dretske, F. (1988) *Explaining Behavior: Reasons in a World of Causes,* Massachusetts Institute of Technology, Cambridge Massachusetts.

Dubpernell, G. (1978), 'Evidence of the Use of Primitive Batteries in Antiquity', in: Dubpernell, G. and Westbrook, J. H. (eds.), *Selected Topics in the History of Electrochemistry,* The Electrochemical Society, Princeton NJ: 1-22.

Eggert, G. (1996) 'The Enigmatic 'Battery of Baghdad'- Scientific Theories on the Ancient Uses of a 2,000 Year Old Finding', in: *Skeptical Inquirer* 20 (3): 31-34.

Elton, M. (2003) *Daniel Dennett; Reconciling Science and Our Self-Conception,* Polity Press, Cambridge.

Enç, B. (1982) 'Intentional States of Mechanical Devices', in: *Mind* XCI: 161-182.

Feenberg, A. (1999) *Questioning Technology,* Routledge Press, London.

Fodor, J. (1974), 'Special Sciences', in: *The Language of Thought,* Harvard University Press, Cambridge MA: 127-145.

----- (1980) 'Methodological Solipsism Considered as a Research Strategy in Cognitive Science', in: *Behavioral and Brain Sciences* 3: 63-109.

----- (1987/1988) *Psychosemantics; The Problem of Meaning in the Philosophy of Mind,* MIT Press, Cambridge MA.

Gasper, P. (1992) 'Reduction and Instrumentalism in Genetics', in: *Philosophy of Science* 59: 655-670.

Gelman, S. A. and Bloom, P. (2000) 'Young Children are Sensitive to How an Object was Created When Deciding What to Name it', in: *Cognition* 76: 91-103.

Gergely, G. and Csibra, G. (2003) *Teleological Reasoning in Infancy: The Naive Theory of Rational Action,* Netherlands, Elsevier Science.

Gergely, G., Nadasdy, Z., Csibra, G., and Biro, S. (1995) 'Taking the Intentional Stance at 12 Months of Age', in: *Cognition* 56: 165-193.

German, T. and Barrett, H. C. (2005) 'Functional Fixedness in a Technologically Sparse Culture', in: *Psychological Science* 16 (1): 1-5.

German, T. and Johnson, S. (2002) 'Function and the Origins of the Design Stance', in: *Journal of Cognition and Development* 3: 279-300.

Gerrans, P. and Kennett, J. (eds.) (2006) 'Empirical Research and the Nature of Moral Judgement', *Philosophical Explorations* 9 (1).

Gibson, K. R. and Ingold, T. (1994) *Tools, Language and Cognition in Human Evolution,* Cambridge University Press, Cambridge, UK.

Godfrey-Smith, P. (2003) 'Folk Psychology Under Stress: Comments on Susan Hurley's 'Animal Action in the Space of Reasons'', in: *Mind and Language* 18 (3): 266-272.

Gould, S. J. and Lewontin, R. (1979) 'The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme', in: *Proceedings of the Royal Society* B205: 581-598.

Griffin, D. R. (1976) *The Question of Animal Awareness: Evolutionary Continuity of Mental Experiences,* Rockefeller University Press, New York.

Griffin, R. and Baron-Cohen, S. (2002), 'The Intentional Stance: Developmental and Neurocognitive Perspectives', in: Brook, A. and Ross, D. (eds.), *Daniel Dennett,* Cambridge University Press, Cambridge, UK: 83-116.

Hampton, J. A. (1995) 'Testing the Prototype Theory of Concepts', in: *Journal of Memory and Language* 34: 686-708.

Haugeland, J. (1998) *Having Thought,* Harvard University Press, Cambridge MA/London.

Heidegger, M. (1977) *The Question Concerning Technology,* Harper and Row, New York.

Hurley, S. (2003a) 'Animal Action in the Space of Reasons', in: *Mind and Language* 18 (3): 231-257.

----- (2003b) 'Making Sense of Animals: Interpretation vs. Architecture', in: *Mind and Language* 18 (3: 273-280).

Hurley, S. and Nudds, M. (eds.)(2006) *Rational Animals?,* Oxford University Press, London.

Jackson, F. and Pettit, P. (2004), 'Program Explanation; A General Perspective', in: Jackson, F., Pettit, P., and Smith, M. (eds.), *Mind, Morality and Explanation; Selected Collaborations,* New York, Clarendon Press: 119-130.

Jansen, W., Fickenfrerichs, H., Peper, R., and Flintjer, B. (1993) 'Die Batterie der Parther und das Vergolden der Bagdader Gotdschmiede', in: *Chemie in Labor und Biotechnik* 44 (3): 128-133.

Jaswal, V. K. (2006) 'Preschoolers Favor the Creator's Label When Reasoning About an Artifact's Function', in: *Cognition* 99 (3): 83-92.

Jonas, H. (1979/1984) *The Imperative of Responsibility: In Search of an Ethics for the Technological Age,* University of Chicago Press, Chicago.

Keil, F. (1995), 'The growth of causal understandings of natural kinds', in: Sperber, D., Premack, D., and Premack, A. (eds.), *Causal cognition: A multidisciplinary debate,* Oxford University Press, Oxford: 234-262.

----- (1989) *Concepts, Kinds, and Cognitive Development,* MIT Press, Cambridge, Massachusetts.

Kelemen, D. (1999a), 'Beliefs about Purpose: on the Origins of Teleological Thought', in: Corballis, M. and Lea, S. E. G. (eds.), *The Descent of Mind; Psychological Perspectives on Human Evolution*, Oxford University Press, New York: 278-294.

----- (1999b) 'The Scope of Teleological Thinking in Preschool Children', in: *Cognition* 70: 241-272.

----- (1999c) 'Why Are Rocks Pointy? Children's Preference for Teleological Explanations of the Natural World', in: *Developmental Psychology* 35 (6): 1440-1452.

----- (2004) 'Are Children "Intuitive Theists"?: Reasoning about Purpose and Design in Nature', in: *Psychological Science* 15 (5): 295-301.

Kemler Nelson, D. G. (2004) 'Two- and three-year-olds infer and reason about design intentions in order to categorize broken objects', in: *Developmental Science* 7 (5): 543-549.

Kemler Nelson, D. G., Herron, L., and Morris, C. (2002) 'How Children and Adults Name Broken Objects: Inferences and Reasoning About Design Intentions in the Categorization of Artifacts', in: *Journal of Cognition and Development* 3 (3): 301-332.

Kim, J. (1989) 'The Myth of Nonreductive Materialism', in: *Proceedings and Addresses of the American Philosophical Association* 63 (3): 31-47.

----- (1993), 'The Non-reductivist's Troubles with Mental Causation', in: Heil, J. and Mele, A. R. (eds.), *Mental Causation*, Clarendon Press, Oxford: 189-210.

----- (1998) *Mind in a Physical World; An Essay on the Mind-Body Problem and Mental Causation,* MIT Press, Cambridge, MA.

----- (2002) 'The Layered Model: Metaphysical Considerations', in: *Philosophical Explorations* V (I): 2-20.

Kiraly, I., Jovanovic, B., Prinz, W., Aschersleben, G., and Gergely, G. (2003) 'The Early Origins of Goal Attribution in Infancy', in: *Consciousness and Cognition* 12 (4): 752-769.

Kitcher, P. (1984) '1953 and All That. A Tale of Two Sciences', in: *The Philosophical Review* 93 (3): 335-373.

----- (1985) 'Two Approaches to Explanation', in: *Journal of Philosophy* : 632-639.

Knobe, J. (2003) 'Intentional action in folk psychology: an experimental investigation', in: *Philosophical Psychology* .

König, W. (1938) 'Ein Galvanisches Element aus der Partherzeit?', in: *Forschungen und Forschritte* 14 (1): 8-9.

Kornblith, H. (ed.)(1988) *Naturalizing Epistemology,* MIT Press, Cambridge, MA.

----- (2002) *Knowledge and Its Place in Nature,* Clarendon Press, Oxford.

Kroes, P. and Meijers, A. W. M. (eds.) (2006) 'Special Issue: The Dual Nature of Artefacts', *Studies in the History and Philosophy of Science* 37 (1).

Landau, B., Smith, L., and Jones, S. (1998) 'Object Shape, Object Function and Object Name', in: *Journal of Memory and Language* 38: 1-27.

Latour, B. (1992), 'Where Are The Missing Masses? The Sociology of a Few Mundane Artifacts', in: Bijker, W. and Law, J. E. (eds.), *Shaping Technology, Building Society. Studies in Sociotechnical Change,* MIT Press, Cambridge/London: 225-258.

Lewens, T. (2002) 'Adaptationism and Engineering', in: *Biology and Philosophy* 17 (1: 1-31).

----- (2004) *Organisms and Artifacts; Design in Nature and Elsewhere,* MIT Press, Cambridge MA.

Lycan, W. G. (2004/2000) *Philosophy of Language,* Routledge, New York/London.

Malt, B. C. and Johnson, E. C. (1992) 'Do Artifact Concepts have Cores?', in: *Journal of Memory and Language* 31: 195-217.

Margolis, E. and Laurence, S. (eds.)(2007) *Creations of the Mind; Theories of Artifacts and their Representation,* Oxford University Press, Oxford.

Matan, A. and Carey, S. (2001) 'Developmental Changes Within the Core of Artifact Functions', in: *Cognition* 78: 1-26.

McCauley, R. N. (1996), 'Explanatory Pluralism and the Co-Evolution in Science', in: McCauley, R. N. (ed.), *The Churchlands and their Critics,* Basic Blackwell, Oxford: 17-47.

McDowell, J. (1994) *Mind and World,* Harvard University Press, Cambridge.

McLaughlin, P. (2001) *What Functions Explain; Functional Explanation and Self-Reproducing Systems,* Cambridge University Press, Cambridge.

Mele, A. R. (ed.)(1997) *The Philosophy of Action,* Oxford University Press, Oxford.

Meyering, T. (2000) 'Physicalism and Downward Causation in Psychology and the Special Sciences', in: *Inquiry* 43: 181-202.

Millgram, E. (ed.)(2001) *Varieties of Practical Reasoning,* Massachusetts Institute of Technology, Cambridge, MA.

Millikan, R. G. (1991) *Language, Thought, and Other Biological Categories; New Foundations for Realism,* MIT Press, Cambridge, MA/ London, England.

----- (1993) *White Queen Psychology and Other Essays for Alice,* MIT Press, Cambridge.

----- (2000), 'Reading Mother Nature's Mind', in: Ross, D., Brook, A., and Thompson, D. (eds.), *Dennett's Philosophy; A Comprehensive Assessment,* MIT Press, Cambridge/London: 55-75.

Nagel, T. (1974) 'What is it Like to be a Bat?', in: *Philosophical Review* 4: 435-450.

Ortiz de Montellano and , B. (1991) 'Multicultural Pseudoscience: Spreading Illiteracy Among Minorties', in: *Skeptical Inquirer* 16 (1): 46-50.

Paszthory, E. (1989) 'Electricity Generation or Magic? The Analysis of an Unusual Group of Finds from Mesopotamia', in: *MAZCA Research Papers in Science and Archaeology* 6: 31-38.

Pettit, P. (1993) *The Common Mind; An Essay on Psychology, Society and Politics,* Oxford University Press, New York, Oxford.

Preston, B. (1998a) 'Cognition and Tool Use', in: *Mind and Language* 13 (4): 513-547.

----- (1998b) 'Why is a Wing like a Spoon? A Pluralist Theory of Function', in: *Journal of Philosophy* 95: 215-254.

Putnam, H. (1975) *Mind, Language, and Reality: Philosophical Papers,* Cambridge University Press, Cambridge.

Quine, W. v. O. (1960) *Word and Object,* MIT Press, Cambridge.

----- (1970) 'On the Reasons for the Indeterminacy of Translation', in: *Journal of Philosophy* 67: 178-183.

----- (1987) 'Indeterminacy of Translation again', in: *Journal of Philosophy* 84 (1): 5-10.

----- (1988), 'Epistemology Naturalized', in: Kornblith, H. (ed.), *Naturalizing Epistemology*, MIT Press, Cambridge, MA: 15-29.

Ramsey, W., Stich, S. P., and Garon, J. (2001), 'Connectionism, Eliminativism, and the Future of Folk Psychology', in: Greenwood, J. D. (ed.), *The Future of Folk Psychology*, Cambridge University Press, Cambridge: 93-119.

Ratcliffe, M. (2001) 'A Kantian Stance on the Intentional Stance', in: *Biology and Philosophy* 16: 29-52.

Renfrew, C. and Zubrow, E. B. W. (2000) *The Ancient Mind; Elements of Cognitive Archaeology,* Cambridge University Press, Cambridge, UK.

Ristau, C. (1991) *Cognitive Ethology: The Minds of other Animals,* Lawrence Erlbaum Associates, Hillsdale, N.J.

Ritzer, G. (1996) *Classical Sociological Theory,* McGraw-Hill, New York.

Rorty, R. (2001/1982) *Consequences of Pragmatism,* University of Minnesota Press, Minneapolis.

Rosenberg, A. (1997) 'Can Physicalist Antireductionism Compute the Embryo?', in: *Philosophy of Science* 64 (Proceedings): S359-S371.

Ross, D. (2000), 'Rainforest Realism: A Dennettian Theory of Existence', in: Ross, D., Brook, A., and Thompson, D. (eds.), *Dennett's Philosophy; A Comprehensive Assessment*, MIT Press, Cambridge MA: 147-168.

Ross, D., Brook, A., and Thompson, D. (2000) *Dennett's Philosophy; a Comprehensive Assessment,* MIT Press, Cambridge/London.

Ryle, G. (1949/1983) *The Concept of Mind,* Penguin Books, Harmondsworth.

Salmon, M. H. (1982) *Philosophy and Archaeology,* Academic Press, New York.

Schick, K. and Toth, N. (1993) *Making Silent Stones Speak; Human Evolution and the Dawn of Technology,* Simon & Schuster, New York.

Searle, J. (1987) 'Indeterminacy, Empiricism, and First person', in: *Journal of Philosophy* 84 (3): 123-146.

----- (1992) *The Rediscovery of the Mind,* MIT Press, Cambridge.

----- (1996) *The Construction of Social Reality,* Penguin Books.

Sellars, W. (1963) *Science, Perception, and Reality,* Routledge and Kegan, London.

Seyfarth, R. M. and Cheney, D. L. (2002), 'Dennett's Contribution to Research on the Animal Mind', in: Brook, A. and Ross, D. (eds.), *Daniel Dennett,* Cambridge University Press, Cambridge: 117-139.

Shoemaker, S. (1996) *The First-Person Perspective and Other Essays,* Cambridge University Press, Cambridge.

Skinner, B. F. (1938) *The Behavior of Organisms,* Appleton-Century-Crofts, New York.

Sloman, S. A. and Malt, B. C. (2003) 'Artifacts are not Ascribed Essences, nor are they Treated as Belonging to Kinds', in: *Language and Cognitive Processes* 18 (5/6): 563-582.

Slors, M. (1996) 'Why Dennett Cannot Explain What it is to Adopt the Intentional Stance', in: *Philosophical Quarterly* 46 (182): 93-98.

Stein, E. (1996) *Without Good Reason; The Rationality Debate in Philosophy and Cognitive Science,* Clarendon Press, Oxford.

Sterelny, K. (2001) *The Evolution of Agency and Other Essays,* Cambridge University Press, Cambridge, UK.

----- (2003a) 'Charting Control-Space: Comments on Susan Hurley's 'Animal Action in the Space of Reasons'', in: *Mind and Language* 18 (3: 257-265).

----- (2003b) *Thought in a Hostile World; The Evolution of Human Cognition,* Blackwell, Cornwall.

Stich, S. P. (1990) *The Fragmentation of Reason,* Massachusetts Institute of Technology, Cambridge, MA.

----- (1991), 'Scientific Versus Folk Psychology', in: Rosenthal, D. M. (ed.), *The Nature of Mind,* Oxford University Press, New York: 590-600.

Strawson, P. (1968), 'Freedom and Resentment', in: Strawson, P. (ed.), *Studies in the Philosophy of Thought and Action,* Oxford University Press, London: 71-96.

Thagard, P. and Nisbett, R. E. (1983) 'Rationality and Charity', in: *Philosophy of Science* 50: 250-267.

Thomasson, A. (2007) *Ordinary Objects,* Oxford University Press, Oxford.

Tollefsen, D. (2002) 'Challenging Epistemic Individualism', in: *Protosociology* 16: 86-117.

Vaesen, K. (forthcoming) *A Philosophical Essay on Artifacts and Norms*. Dissertation, Technical University of Eindhoven.

Vaesen, K. and Amerongen, M. v. (forthcoming) 'Optimality and Intent; Limitations of Dennett's Artifact Hermeneutics'. *Philosophical Psychology*.

Vermaas, P. E. and Houkes, W. (2003) 'Ascribing Functions to Technical Artefacts: A Challenge to Etiological Accounts of Functions', in: *British Journal for the Philosophy of Science* 54 (2: 261-289).

Viger, C. (2000), 'Where Do Dennett's Stances Stand?', in: Ross, D., Brook, A., and Thompson, D. (eds.), *Dennett's Philosophy*, MIT Press, Cambridge MA: 131-145.

Von Daniken, E. (1993) *Raumfahrt im Altertum,* Bertelsmann, München.

Waters, C. K. (1990) 'Why the Anti-Reductionist Consensus won't Survive. The Case of Classical Mendelian Genetics', in: *PSA 1990* 1: 125-139.

Watson, J. B. (1919) *Psychology from the Standpoint of the Behaviorist,* Lippencott, Philadelphia.

Wellman, H. M. (1990) *The Child's Theory of Mind,* MIT Press, Cambridge, Massachusetts.

Wimmer, H. and Perner, J. (1983) 'Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception', in: *Cognition* 13: 103-128.

Wimsatt, W. and Beardsley, M. (1946/1974), 'The Intentional Fallacy', in: Newton-De Molina (ed.), *On Literary Intention*, Edinburgh University Press, Edinburgh: 1-13.

Index of names

Summary (in English)

The Interpretation of Artifacts; A Critique of Dennett's Design Stance

Technological artifacts are a pervasive part of human life. They are, however, largely ignored in the analytic philosophical tradition, especially by philosophical naturalists. Being mind-dependent phenomena, tied up with human intentionality, analytic philosophers have largely found the topic unscientific, not objective, or simply trivial. An important exception is Daniel Dennett, who puts design at the heart of his naturalistic theory of mind.

Dennett's theory of functions, which I call the optimality account of functions, is an alternative for intentionalistic accounts of functions. The optimality account says that a function of a certain item is what it is best able to do given the circumstances and not what its designer intended it to be. It takes seriously the idea that a pure physical description of technical artifacts is not enough to fully understand them. Optimality is, after all, a normative notion. But by avoiding reference to intentions of designers or users of the artifact in the conceptualization of function, it is an interesting attempt to prevent problems that arise from intentionalistic accounts of technical function.

Dennett defends the optimality account because he needs a notion of biological function that does not refer to intentions in order to prevent circularity. For intentions have, eventually, to be explained by means biological functions and design. Dennett furthermore wants a generic notion of both biological and artifact function, so artifact functions have to be conceptualized without reference to intentions as well.

I have formulated two standards, empirical correctness and methodological value, by means of which I will evaluate the optimality account. The standard of empirical correctness says that a theory of function attribution should match empirical reality (descriptive). The methodological standard says that it should provide a good scientific or practical method (normative). I derive these two standards directly from Dennett's work, so they should certainly be met by a Dennettian theory of artifact interpretation. The two standards are quite general, however. They should be embraced by many philosophers, certainly those working in the more naturalistic and pragmatic tradition.

I argue that the optimality account fails to meet both standards. Cognitive psychological research strongly suggests that the intentionalistic account is the better description of our ordinary way of ascribing functions to artifacts than the optimality account. That is to say, our everyday conceptualization of artifacts is largely driven by what we believe the designer of the artifact intended the artifact to be. The optimality account thus fails the empirical standard.

The optimality account has a weak score on the methodological standard as well. In scientific and practical fields where we would expect Dennett's design stance to work demonstrably better than intentionalistic accounts of function interpretation, it fails to show its worth. First, reasoning about functions yields better results if we do take into account intentions and the intentional context in which they are created and used (even when intentions are hard to trace). Secondly, contra Dennett’s claims, the predictive value of the optimality account is limited. Furthermore, the practical design stance as method is inherently intentional. Contra Dennett, reasoning about functions thus yields better results when we do take intentions into account.

I conclude that Dennett's optimality account of the design stance, promising as it is, fails on its own standards. Worse, it seriously threatens to backfire on his larger theory of the stances. I conclude that the optimality account has limited applications and would certainly improve if it were able to take into account information about the intentional context in which it was created or used and its intended design.

Nederlandstalige samenvatting

In deze samenvatting behandel ik de hoofdlijnen van het proefschrift. Ik permitteer me een zekere vrijheid, want deze samenvatting is niet zozeer bedoeld als een technische samenvatting van wat er in het proefschrift wordt behandeld, maar eerder als een zeer verkorte vogelvlucht die ook voor niet-ingewijden in de filosofie begrijpelijk is.

Introductie

Technologie is onmiskenbaar een wezenlijk en kenmerkend onderdeel van het menselijk leven. Dat maakt het onderwerp voor filosofen interessant. Vooral in de zogenaamde continentale traditie is er van oudsher veel aandacht voor het onderwerp. In de analytische traditie, die vooral beoefend wordt in Engelstalige landen als Groot Brittannië, de Verenigde Staten en Australië, wordt het onderwerp gemeden. Langzamerhand begint de interesse voor techniek ook daar toe te nemen. Het gaat dan vooral om de vraag hoe technische artefacten geconceptualiseerd moeten worden.

Het proefschrift start met een assumptie die ontleend is aan het Dual Nature project: dat de functie van een technisch object (of in filosofisch jargon: 'artefact') bestaat uit twee dimensies: een fysieke en een mentale. De fysieke dimensie gaat dan over de opbouw van het ding, de mentale dimensie over de gedachte erachter: de bedoeling waarvoor het is gemaakt of wordt gebruikt. Beide dimensies zijn volgens deze these noodzakelijk om een technisch artefact te begrijpen. Bijvoorbeeld: een stukje hout is uitsluitend een *tandenstoker* als het (1) fysiek geschikt is om tanden mee te stoken (2) het ontworpen of gebruikt is voor het doel tanden stoken.

Deze conceptualisering is niet zonder problemen. Door artefacten in termen van bedoelingen of intenties te duiden, wat intuïtief klopt, krijgen we automatisch te maken met de vraag wat intenties, of breder: mentale toestanden, zijn. En daarmee krijgt de thematiek van technische artefacten onmiddellijk te maken met een zwaar bediscussieerd probleem in de filosofie: hoe verhouden materie en het denken zich tot elkaar? Artefacten behoren dan tot de categorie 'subjectieve' of 'intersubjectieve' verschijnselen, verschijnselen die hun bestaan danken aan ons denken erover, en als zodanig niet tot de echte wereld behoren, die immers los van onze gedachten bestaat. Daarom wordt vaak geprobeerd technische artefacten te duiden zonder referentie naar intentionele toestanden.

Onder andere om deze reden behandel ik in mijn proefschrift een alternatieve filosofische theorie, die technische artefacten niet conceptualiseert met behulp van intenties, maar in termen van optimaliteit: de Intentionele Systemen Theory (IST) van de filosoof Daniel C. Dennett. Deze theorie gaat vooral over de vraag wat intentionele toestanden precies zijn, en heeft belangrijke implicaties voor de vraag wat functies van technische artefacten zijn.

IST zegt dat we grofweg drie soorten houdingen kunnen onderscheiden. De *fysieke houding* is het meest bekend. Hiermee verklaren we het gedrag van fysieke objecten, min of meer zoals in de natuurwetenschap. Maar niet alle fenomenen zijn zo gemakkelijk te verklaren en te voorspellen met deze fysieke houding. Menselijk gedrag, bijvoorbeeld, is veel te ingewikkeld om in puur fysische termen te begrijpen, zeker in praktische contexten. Toch kunnen we menselijk gedrag behoorlijk goed voorspellen. Dit hebben we te danken aan de *intentionele houding*. De intentionele houding nemen we in ten opzichte van allerlei handelende wezens, ook wel *agents* genoemd (iets is een agent áls we succesvol de intentionele houding tegenover hem kunnen innemen). We schrijven gedachten, wensen en intenties aan een agent toe, nemen aan dat de agent grofweg rationeel zal zijn, en voorspellen en verklaren daarmee zijn handelen. En dit is volgens Dennett precies wat een gedachte (wens, intentie, etc.) is: een toegeschreven toestand, die *handig* is om handelingen mee te verklaren en te begrijpen. Een gedachte is dus eigenlijk een zaak van interpretatie.

Tot slot is er de *ontwerphouding*. Hiermee voorspellen we het gedrag van technische artefacten, maar ook van objecten die zich gedragen als iets dat ontworpen is (waaronder biologische objecten zoals een oog, een hart, een stamper). We nemen aan dat het object ontworpen is en functioneert zoals bedoeld. Daarmee kunnen we het gedrag van het object voorspellen. Een mooi voorbeeld is dat van een lift. Dankzij de ontwerphouding kunnen we zonder iets te weten over de interne mechaniek met voldoende zekerheid voorspellen dat de lift naar de zevende verdieping zal gaan. Als we tenminste op het daartoe bestemde knopje #7 drukken.

De ontwerphouding vervult een belangrijke rol in Dennett's theorie. Niet op de laatste plaats omdat hij meent dat de ontwerphouding evenzeer van toepassing is op technische artefacten als op biologische fenomenen. Want Dennett is een biologisch naturalist: de geest moet volgens Dennett begrepen worden als een biologisch fenomeen, en als zodanig als een door de natuur 'ontworpen' artefact.

Als fervent anti-creationist kan Dennett niet zomaar beweren dat er sprake is van ontwerp in de biologische natuur. Ontwerp veronderstelt immers een ontwerper. Desondanks ontwikkelt Dennett een notie van biologisch ontwerp, die geen ontwerper veronderstelt.

Ik noem dit het optimaliteitsprincipe. Zowel technische als biologische ontwerpen zijn volgens dit account te begrijpen via de notie van optimaliteit en zonder referentie naar intenties. Deze notie van ontwerp vormt de kern van dit proefschrift.

Ik onderzoek de vraag of Dennett's optimaliteitsprincipe, en de bijbehorende theorie van de drie houdingen, een adequate wijze van conceptualisering biedt van functies van technische artefacten. Hiervoor ontwikkel ik twee evaluatieve standaards die ik direct afleid uit Dennett's theorie zelf: empirische adequaatheid en methodologische bruikbaarheid. Het is dus een *interne* evaluatie. Interne evaluaties zijn veel sterker dan externe evaluaties, die per definitie uitgaan van *externe* standaards die de filosoof in kwestie niet hoeft te delen. Dat maakt het leveren van kritiek, en het verdedigen ertegen, relatief gemakkelijk. In het geval van Dennett is dit speciaal van belang, aangezien Dennett's ideeën over de geest en biologie gegrond zijn in zulke eigenzinnige opvattingen, dat een externe discussie direct uitmondt in een discussie over ultieme filosofische grondbeginselen.

De opbouw van het argument is als volgt. De eerste helft van het proefschrift is een voorbereiding voor de tweede helft. Hier bespreek ik achtereenvolgens globaal de discussie over het interpreteren van attitudes (hoofdstuk 2), de basisbeginselen van Dennett's theorie (hoofdstuk 3), en de filosofische plaatsing van Dennett's theorie over attitudes. De eerste helft is vooral interpretatie van Dennett's theorie, de tweede helft bevat de eigenlijke argumentatie:

(1) uitwerking van het optimaliteitsprincipe (hoofdstuk 5)

(2) toetsing van het optimaliteitsprincipe aan de hand van de eerste standaard (hoofdstuk 6)

(3) toetsing van het optimaliteitsprincipe aan de hand van de tweede standaard (hoofdstuk 7)

Waarna vervolgens een conclusie (hoofdstuk 8) getrokken kan worden.

2. Filosofie van attitudes

Dennett's theorie over technische artefacten en ontwerpen, en veel andere filosofische theorieën over functies, begint bij de filosofie van attitudes. 'Attitude' is een filosofisch-technische term, die verwijst naar mentale toestanden als gedachten, wensen, gevoelens, intenties of bedoelingen. Deze worden ook wel aangeduid met intentionele toestanden, propositionele toestanden of psychologische toestanden. 'Pietje denkt dat het bier op is', is een voorbeeld van het toeschrijven van een denk-attitude met de inhoud 'het bier is op' aan Pietje. Attitudes zijn voor filosofen interessant, omdat ze zo'n belangrijke rol spelen in het alledaags leven, maar zo lastig in een wetenschappelijk kader te plaatsen zijn. We schrijven ze immers te pas en te onpas toe aan anderen en beperken ons daarbij niet alleen tot mensen. Het gebruik van attitude-termen wordt dan ook wel volkspsychologie (*Folk Psychology*) genoemd. Dennett's theorie over de drie houdingen kan niet goed begrepen worden zonder iets te weten over het ingewikkelde filosofische debat dat over deze attitudes gevoerd wordt. Gezien de grote hoeveelheid stromingen en terminologische nuances in dit debat ligt Babylonische spraakverwarring op de loer. De rol van dit hoofdstuk is om de verschillende mogelijke benaderingen in elk geval globaal helder te krijgen, zodat Dennett's ideeën erover beter geplaatst kunnen worden.

Grofweg zijn er drie belangrijke, met elkaar verweven, onderscheidingen in deze discussie. De eerste onderscheiding is tussen attitudes en het toeschrijven van attitudes. Dat we attitudes toeschrijven, staat buiten kijf. Maar waarnaar verwijzen deze toeschrijvingen? Dit is belangrijk, aangezien toeschrijvingen verkeerd zouden kunnen zijn. In stripboeken worden gedachten en wensen doorgaans aangeduid met spraak- en denkwolkjes, die suggereren dat attitudes zich ergens in de hersenen of geest bevinden. Maar of dat zo is (en hoe dan) is een vraag die zich niet gemakkelijk laat beantwoorden. Waar in de hersenen bevindt zich bijvoorbeeld inhoud van de attitude 'het bier is op'? Dennett's positie is dat er niets is behalve attitude *toeschrijving*. Attitudes zelf bestaan niet, althans niet op de manier waarop wij ze beschrijven, gebruiken en toeschrijven.

Het tweede onderscheid is tussen descriptieve en normatieve benaderingen. Descriptieve benaderingen richten zich op vragen als 'hoe schrijven we *de facto* attitudes toe?' en 'wat zijn attitudes nu eigenlijk'? Normatieve benaderingen gaan over de vraag hoe we attitudes *zouden moeten* toeschrijven. Is ons alledaags taalgebruik voor correctie vatbaar? Voor Dennett vallen beide samen. Een goede toeschrijving is een nuttige toeschrijving, een toeschrijving die ons helpt handelingen te voorspellen. En aangezien onze alledaagse toeschrijvingen zeer nuttig zijn (menselijk leven zou niet mogelijk zijn zonder attitude toeschrijving), zijn onze dagelijkse toeschrijvingen legitiem.

Het laatste onderscheid is tussen alledaagse attitudes en wetenschappelijke attitudes. Hierbij speelt de vraag of attitudes via het alledaags taalgebruik geconceptualiseerd dienen te worden, of dat we op zoek moeten naar een wetenschappelijk begrip ervan. Attitudes zijn alledaagse begrippen en daarom ligt het voor de hand ze ook als zodanig te conceptualiseren. In dat geval onderzoeken we hoe attitude-taal daadwerkelijk gebruikt wordt, en corrigeren we in geval van contradicties en inconsistenties. Dit is een controversiële methode, waarvan niet altijd duidelijk is wat hij oplevert. Attitude toeschrijving heeft zo'n voorspellende kracht, dat het voor de hand ligt te denken dat attitudes op enigerlei wijze 'echt' bestaan. In dat geval zou ze ook wetenschappelijk te onderzoeken moeten zijn. En als je attitudes kunt onderzoeken en hun bestaan kan worden aangetoond, zou dat ons alledaags taalgebruik erover legitimeren en de kracht ervan verklaren. Het is echter onwaarschijnlijk dat het wetenschappelijk concept van een attitude zal overeenkomen met het alledaagse concept. Mocht er een wetenschappelijk begrip van attitudes ontwikkeld worden, dan is het zeer de vraag of dat nog iets te maken heeft met het alledaagse begrip ervan. Dennett tracht door deze discussie heen te breken door te stellen dat een wetenschappelijk *begrip* van attitudes helemaal niet nodig is (attitudes bestaan immers niet als zodanig), maar dat de alledaagse *methode* van attitude toeschrijving wel wetenschappelijk waardevol kan zijn.

3. De essentie van Dennett's filosofie

In hoofdstuk 3 wordt Dennett's algemene filosofie van de geest verder uitgewerkt, alsook de onderliggende principes ervan. Op basis hiervan kan ik de twee standaarden formuleren op basis waarvan zijn optimaliteitsaccount kan worden beoordeeld.

Dennett's theorie over de geest omvat een zeer breed spectrum van onderwerpen zoals bewustzijn, vrijheid, evolutie en intelligentie. Zijn theorie over attitudes is daarvan een klein maar belangrijk onderdeel. Maar de brede theorie kan gekarakteriseerd met twee termen: naturalisme en pragmatisme. Ik interpreteer Dennett als een pragmatist die vindt dat wetenschap de meest werkbare manier is om vragen over de geest te beantwoorden. Een cruciaal derde element is Dennett's interpretationisme.

Interpretationisme is een filosofische stroming gebaseerd op Quines these van radicale vertaling (*thesis of radical translation*) en zegt dat we altijd interpreteren als we attitudes toeschrijven. Dit heeft een aantal belangrijke consequenties voor ons begrip van attitudes. Op de eerste plaats kan het, omdat attitudes volgens de interpretationist noodzakelijkerwijs door de interpretator worden toegeschreven, soms onmogelijk zijn een exacte bepaling te geven van de inhoud van een attitude (onbepaaldheid). Daarnaast worden toeschrijvingen gereguleerd middels het principe van welwillendheid (*principle of charity*), wat wil zeggen dat we bij het interpreteren moeten aannemen dat de attitudes die we toeschrijven consistent met elkaar zijn en samenhangen. Tenslotte kan er nooit sprake zijn van geïsoleerde toeschrijving van een attitude, maar dient een attitude altijd geïnterpreteerd te worden in relatie tot een netwerk van andere attitudes (betekenisholisme / *meaning holism*).

Dennett staat het meest bekend als naturalist. Filosofische vragen over de geest vinden, volgens Dennett, een eerste antwoord in wetenschap. En er is niets van de geest dat wetenschappelijk niet verklaarbaar of onderzoekbaar is. Elke toeschrijvingstheorie dient daarnaast voldoende gegrond te zijn in accurate beschrijvingen van de realiteit die door wetenschappelijk onderzoek kenbaar is.

Dennett's pragmatisme komt duidelijk tot uiting in zijn theorie over attitudes. Het enige criterium op basis waarvan we kunnen vaststellen of een toeschrijving van een attitude juist is of niet, is een pragmatisch criterium: een toeschrijving is juist als deze voor de toeschrijver zinvol of nuttig is. Dennett geeft hier snel een naturalistische wending aan door aan te geven dat vooral toeschrijvingen met veel voorspellende kracht zinvol zijn. Wetenschappelijke toeschrijvingen zijn dit natuurlijk bij uitstek.

Op basis hiervan kom ik tot twee standaarden waaraan een Dennetiaanse conceptualisering (dus toeschrijvingswijze) moet voldoen. Uiteraard moet het passen binnen een interpretationistisch kader, maar daarnaast:

(1) dient deze aan te sluiten bij de empirische realiteit (empirische adequaatheid)

(2) dient deze een bruikbare methodologie te zijn, wetenschappelijk of anderszins (methodologische bruikbaarheid)

Deze standaards zijn voor veel filosofen belangrijk, zeker voor naturalisten, maar ze zijn cruciaal voor een filosofische theorie als die van Dennett.

4. Interpretationisme als non-reductionisme

Voordat Dennett's optimaliteits account nader kan worden uitgewerkt en vervolgens geëvalueerd aan de hand van de twee criteria, geef ik een nadere (vrij technische) positionering van Dennett's interpretationistische theorie. Het maakt voor de verdere interpretatie namelijk veel uit of Dennett's interpretationisme als een ontologische theorie gezien wordt, of een epistemologische.

Ontologie gaat over 'wat er is' en de vraag waar de wereld nu eigenlijk uit bestaat. Iedereen die beweert dat attitudes (of meer in het algemeen: de geest) bestaan, krijgt te maken met de vraag hoe die attitudes zich dan verhouden tot de fysieke realiteit. Bestaan attitudes uiteindelijk louter uit fysische elementen? Zo ja: dan bestaan attitudes dus eigenlijk niet en ben je een reductionist (het mentale kan herleid worden tot het fysieke). Of bestaat er naast die fysieke realiteit nog iets anders? En zo ja, hoe serieus moeten we dat 'iets' dan nemen? Hoe kan het bijvoorbeeld invloed uitoefenen in en op de fysieke wereld? Vanuit ontologisch perspectief bestaan mentale toestanden ofwel geheel niet, of zijn het bijzonder mysterieuze entiteiten. Beide zijn geen aantrekkelijke opties voor wie mentale toestanden een serieuze plek wil geven in de wereld.

Veel filosofen plaatsen Dennett in het ontologie-debat en concluderen dat Dennett simpelweg een reductionist is: er is alleen materie en alles wat verwijst naar mentale zaken is louter een manier van spreken. Ik laat zien dat Dennett hiermee tekort wordt gedaan en veel beter in het epistemologische debat geplaatst kan worden. En dan blijkt dat Dennett allesbehalve een reductionist is, maar juist een van de sterkste non-reductionistische posities inneemt.

Epistemologie gaat over de manier waarop we de wereld kennen. Veel wetenschapsfilosofen hanteren een visie waarbij de wereld gezien wordt als uiteindelijk bestaand uit fysieke deeltjes (de uiteindelijke ontologie), maar deze wereld wel op meerdere manieren beschreven kan worden (de epistemologie). Ontologisch gezien kan iemand dus een reductionist zijn, maar epistemologisch een non-reductionist.

Epistemisch non-reductionisme komt voor in diverse sterktes, waarvan ik de meest onderscheidende types behandel. De sterkste variant zegt dat psychologische beschrijvingen (of meer in het algemeen: hogere orde beschrijvingen en verklaringen) een echt ander *type* verklaring zijn dan fysische en daarom onreduceerbaar. Een veel gebruikt voorbeeld in dit verband zijn functionele verklaringen. Neem het hart. We kunnen de werking van het hart in causaal-fysische termen beschrijven en verklaren. Of we kunnen de werking van het hart verklaren in functionele termen: het hart werkt zoals het werkt *om* bloed door het lichaam te pompen. Een ander voorbeeld is het gebruik van generaliserende functionele terminologie, zoals 'voortplantingsgedrag'. Dergelijke generaliserende termen spelen een zelfstandige verklarende rol, waarbij het niet zozeer gaat om de verschillende fysieke instanties van (bijvoorbeeld) voortplantingsgedrag, maar juist om wat al deze instanties in functionele zin gemeenschappelijk hebben. Filosofen spreken in dit verband vaak van 'meervoudige realisatie'. Zoals 'vijf euro' op meerdere manieren fysiek gerealiseerd kan zijn (op een pinpas, in de vorm van een biljet of munten, etcetera), zouden dezelfde psychologische toestanden ook op meerdere manieren gerealiseerd kunnen worden. Veel filosofen, waaronder Dennett, zien intentionele verklaringen als een type functionele verklaring.

Vanuit epistemologisch perspectief heeft Dennett een sterk non-reductionistische psychologische theorie. Ook claimt hij dat intentionele verklaringen van een andere orde zijn dan fysische (vergelijk het verschil tussen de intentionele houding en de fysieke houding). En dat hun kracht juist ligt in hun vermogen te generaliseren. Hetzelfde geldt voor de ontwerp-houding.

Bezien vanuit ontologisch perspectief zijn intentionele verklaringen 'slechts' heuristiek, een manier van praten, onbestaand in de werkelijke wereld. Vanuit epistemologisch perspectief zijn het eigenstandige, en als zodanig waardevolle, typen verklaringen, die ons in staat stellen onderscheidingen aan te brengen in de wereld die fysieke theorieën niet kunnen bieden.

5. Het optimaliteitsprincipe

De aandacht is tot nu toe uitgegaan naar de intentionele houding en intentionele verklaringen. Dat ligt voor de hand, want de intentionele houding vormt de kern van Dennett's filosofie van de geest. De intentionele houding is op zijn beurt weer sterk verbonden met de ontwerphouding.

Net als de intentionele houding is de ontwerphouding in sterke mate idealiserend: als we een handeling interpreteren gaan we ervan uit dat de agent rationeel is; als we een ontwerp interpreteren, gaan we ervan uit dat het optimaal ontworpen is. Dankzij deze aanname winnen zowel ontwerphouding als intentionele houding aan voorspellende kracht. Beide houdingen zijn ook 'constitutief': een agent is een agent als we hem succesvol kunnen interpreteren als een agent (de intentionele houding succesvol kunnen toepassen); een ontwerp is een ontwerp als we het succesvol kunnen interpreteren als een ontwerp. Ook hieruit blijkt weer dat deze houdingen echt van een ander soort zijn dan de fysische houding. De intentionele houding en ontwerphouding geven minder accurate voorspellingen dan de fysieke, maar ze zijn wel sneller in complexe situaties en daardoor meestal efficiënter.

De ontwerphouding wordt toegepast op ontwerpen. Het gaat Dennett dan in eerste instantie om biologische ontwerpen, maar volgens Dennett zijn er geen wezenlijke verschillen tussen het interpreteren van technische ontwerpen (ofwel 'artefacten') en het interpreteren van biologische ontwerpen. Het proefschrift gaat over technische ontwerpen, maar om Dennett's ideeën hierover duidelijk te krijgen, zal dus ook gekeken moeten worden naar zijn ideeën over biologische ontwerpen.

Wat zegt Dennett over (technische) ontwerpen? Ik onderscheid vier stellingen, waarvan de eerste drie helpen om nummer vier, die gaat over het optimaliteitsprincipe, concreet invulling te geven.

(1) these van liberale attitude toeschrijving

Deze these zegt dat we elk artefact dat zich als agent gedraagt, we als agent mogen interpreteren. Dat iets een functioneel ontwerp is, wil niet zeggen dat intentionele interpretatie niet meer mogelijk is. Dennett's klassieke voorbeeld is de thermostaat, die geïnterpreteerd kan worden als een agent met 'gedachten' over de actuele kamertemperatuur, 'wensen' over de ideale kamertemperatuur en op basis van deze gedachten en wensen handelt (de thermostaat 'vertelt' de CV installatie het water op te warmen of juist niet). Deze these is vooral van belang om te laten zien dat Dennett er een zeer liberale theorie van attitudes op nahoudt. Het is bovendien de meest centrale en meest bediscussieerde these in relatie tot zijn Intentionele Systemen Theorie.

(2) these van primaat van ontwerp

Dennett's liberale attitude theorie is uiteindelijk gegrond in de naturalistische these dat alles wat met de geest te maken heeft, een natuurlijk verschijnsel is. De geest is, net als een arm of een oog, een product van de natuurlijke evolutie. Het evolutieproces kan daarbij volgens Dennett uitstekend als een ontwerpproces gezien worden. De geest kan dus ook gezien worden als een ontwerp (en de gewoonte om attitudes toe te schrijven aan anderen uiteraard ook). De implicatie hiervan is cruciaal: de ontwerphouding verklaart de geest en daarmee ook de intentionele houding. Net zoals wij, mensen, technische artefacten kunnen maken die in zekere zin 'denken' en 'willen' (zoals een thermostaat), creëert de natuur 'menselijke artefacten' met gedachten en wensen. Volgens Dennett is er geen essentieel verschil tussen deze twee processen. Onze intentionaliteit verschilt niet wezenlijk van die van een thermostaat, maar is wel oneindig veel interessanter en complexer!

(3) these van functie generaliteit

Zoals eerder opgemerkt verdedigt Dennett een generiek functiebegrip: we schrijven op dezelfde wijze functies toe aan technische als aan biologische artefacten. Veel naturalisten prefereren een generiek functiebegrip. Immers, het is moeilijk houdbaar om te beweren dat technische functies op geheel andere wijze geïnterpreteerd moeten worden dan biologische functies, terwijl techniek uiteindelijk wél te herleiden moet zijn tot biologie.

Aanvankelijk werkte Dennett dit generieke functiebegrip uit door biologische-functionele verklaringen gelijk te stellen aan technisch-functionele verklaringen. Net zoals we een ingenieur kunnen vragen wat hij met een ontwerp bedoeld had, kunnen we dat aan 'Moeder Natuur' vragen. Bij het interpreteren van ontwerpen gaan we er dan vanuit dat de ontwerper het best mogelijke ontwerp heeft gemaakt. In de evolutiebiologie heet een dergelijke vorm van functie-interpretatie *adaptationisme*.

In eerste instantie hanteerde Dennett dus klaarblijkelijk een *intentioneel* generiek functiebegrip, waarbij functies gezien worden als het resultaat van een intentie. Ken de intentie, ken de functie. In reactie op critici heeft Dennett dit begrip later aangepast. Niet alleen blijkt dat het 'bevragen' van de intenties die 'Moeder Natuur' vooral metaforisch moet worden opgevat, ook blijkt dat Dennett af wil van elk functiebegrip waarbij naar intenties verwezen wordt. De reden daarvoor ligt in these 2: als intenties uiteindelijk verklaard kunnen worden in termen van hun ontwerp, dan kunnen ontwerpen niet op hun beurt verklaard worden in termen van intenties. Dat zou de verklaring immers circulair maken.

(4) these van optimaliteit

Opvallend genoeg werkt Dennett zijn these dat ontwerpen niet in termen van intenties geïnterpreteerd moeten worden juist uit in zijn werk over *technische* artefacten. Vooropgesteld moet worden dat Dennett's alternatief, het optimaliteitsprincipe, vooral een negatief account is, waarin hij ageert tegen intentionele functiebegrippen. Het alternatief wordt tamelijk summier uitgewerkt.

Wat zijn Dennett's argumenten tegen een intentioneel functiebegrip van technische artefacten? Dennett geeft de volgende argumenten:

(1) intenties zeggen bar weinig over de functie van een artefact. Dit is vooral een argument tegen wat ik extreem intentionalisme noem. Een extreem intentionalist stelt dat de functie van een artefact volkomen te herleiden is tot de functie die de ontwerper erbij bedacht had. Dit is een onwaarschijnlijke theorie, vindt Dennett, omdat functies van een artefact kunnen veranderen. Dat maakt intentionalisme een slechte voorspellende theorie. De historische functie van een object, vindt Dennett, is niet relevant.

(2) een intentioneel functiebegrip zou te vaag en onbepaald zijn. Dit argument is gericht tegen elk intentioneel functiebegrip, ook één waarbij bijvoorbeeld gebruikersintenties een rol zouden spelen.

Beide argumenten zijn zwak. Het eerste richt zich op een stromanpositie die in werkelijkheid nauwelijks bestaat. En met het tweede argument ondergraaft Dennett zijn eigen theorie: als onbepaaldheid van intentietoeschrijving een probleem zou zijn bij het toeschrijven van functies aan een artefact, dan zou dat ook een probleem zijn voor het toeschrijven van intenties aan agenten. Dennett ontkent, zeer overtuigend, het laatste. Zijn hele theorie van attitude toeschrijving is zelfs gebaseerd op de stelling dat de onbepaaldheid van attitudes (waaronder intenties), ons er niet van moet weerhouden intenties toe te schrijven. Accuraatheid werd immers ingewisseld voor efficiëntie.

Dennett's alternatief is de optimaliteits benadering. Als we een functie interpreteren, dienen we ervan uit te gaan dat het object optimaal ontworpen is. De functie die het object het best kan vervullen (in huidige condities, gegeven de context en gegeven de mogelijkheden die het object fysiek te bieden heet), *is* de functie die het object heeft. Ongeacht de bedoelingen van ontwerpers of gebruikers.

Intentie volgt uit optimaliteit, in plaats van andersom. Als iets bijvoorbeeld een optimale kersen-ontpitter blijkt, dan leiden we *daaruit* af dat dat dus ook de intentie achter het ontwerp moet zijn geweest. Daarmee volgt Dennett de redeneertrant van de 'intentionele drogreden' (*intentional fallacy*) in de literatuurkritiek.

Het beste voorbeeld van het soort optimaliteitsredeneringen waar Dennet op doelt zien we terug in 'reverse engineering'. Bij het ontleden van het ontwerp van een product van de concurrent, tracht de 'reverse engineer' de rationale achter het ontwerp te vinden. Dit doet hij door vanaf het begin af aan aan te nemen dat elk onderdeel is aangebracht met een reden, en geen onderdeel overbodig is. Zo kan de functie van het object, en zijn onderdelen, achterhaald worden.

Concluderend: het optimaliteitsaccount volgt logisch uit Dennett's ideeën over de interpretatie van ontwerpen, maar kan het de toets der eigen kritiek doorstaan? Voldoet het aan de twee standaarden empirische adequaatheid en methodologische bruikbaarheid?

6. De alledaagse ontwerphouding

De kracht van Dennett's theorie over de intentionele houding ligt in de sterke combinatie van enerzijds dichtbij de empirische realiteit blijven, anderzijds daar zo sterk mogelijke normatieve conclusies uit trekken. Zoals vermeld in hoofdstuk 2 is het niet eenvoudig om de vraag te beantwoorden wat de 'juiste' manier van attitude toeschrijving is. Juist naturalisten worden geconfronteerd met het probleem van het overbruggen van de kloof tussen feit en waarde. Uit feiten kunnen immers geen waarden worden afgeleid, maar de naturalist heeft weinig andere fundamenten dan de empirische feiten om op voort te bouwen. Door de intentionele houding als alledaagse methode te duiden, en de praktische methodologische waarde van voorspellende kracht voorop te stellen, lukt het Dennett om empirische realiteit en een waardering daarvan, aan elkaar te koppelen. Een mooi samenspel dus tussen naturalisme en pragmatisme. Juist hierom zijn de twee standaarden die ik geformuleerd heb, empirische correctheid en methodologische bruikbaarheid, zo cruciaal.

In hoeverre is de ontwerphouding (volgens het optimaliteitsprincipe) een adequate weergave van de empirische realiteit? Om deze vraag te beantwoorden is het interessant te kijken naar het cognitief psychologisch onderzoek dat recent wordt uitgevoerd naar de ontwerphouding. Naar de manier waarop mensen (meestal kleine kinderen) artefacten interpreteren en classificeren, om precies te zijn.

In de cognitieve psychologie wordt veel onderzoek gedaan naar de manier waarop mensen typen objecten classificeren. De nadruk ligt hierbij op de classificatie van natuurlijke objecten en causaliteit en vooral op het vermogen te redeneren over de mentale toestand van een ander. Rond de leeftijd van vier of vijf, zijn kinderen meestal redelijk in staat om zich te verplaatsen in een ander, en op basis daarvan zijn gedrag te voorspellen. Zij hebben dan een zogeheten *Theory of Mind*. In Dennettiaanse terminologie kunnen we ook zeggen dat ze volwaardig gebruikers zijn van de intentionele houding en deze kunnen toepassen op anderen.

Aan de *Theory of Mind* gaat een fase vooraf, die voor mijn doelen interessanter is dan de *Theory of Mind* zelf. Al op zeer jonge leeftijd, misschien al vanaf 12 maanden, lijken kinderen onderscheid te maken tussen objecten die ze in causale termen duiden en objecten die ze op intentionele manier begrijpen: tussen objecten en agenten. Als kinderen een object te zien krijgen dat zich gedraagt als een agent (bijvoorbeeld een balletje op een computerscherm dat over een muurtje heen 'springt'), dan verwachten ze van zo'n type object ander gedrag dan van normale fysische objecten (een balletje dat naar een muur rolt en daar blijft liggen). Van het agent-achtige balletje verwachten ze *rationeel* gedrag. Als het agent-achtige balletje irrationeel gedrag vertoont, bijvoorbeeld een sprongetje maakt terwijl dat niet nodig is, staren kinderen veel langer naar het scherm dan normaal. De cruciale veronderstelling bij dit type onderzoek is dat kinderen langer staren naar het scherm als de gebeurtenissen niet overeenkomen met hun verwachtingen.

Als dit onderzoek klopt, zou het Dennett's ideeën over de intentionele houding in empirisch opzicht verder versterken. Het zou immers betekenen dat kinderen vanaf zeer jonge leeftijd al gedachten, wensen en doelen aan abstracte objecten kunnen toekennen en op basis daarvan gedrag voorspellen. Het belangrijke verschil met de *Theory of Mind* die later ontstaat zou zijn dat kinderen op zeer jonge leeftijd nog geen *foute* overtuigingen kunnen toeschrijven. En dit zou betekenen dat je geen *Theory of Mind* nodig hebt om intenties te kunnen toeschrijven: het toeschrijven van een gedachte, veronderstelt geen *concept* van een gedachte.

Het uitstapje naar de primitieve intentionele houding illustreert hoe empirisch onderzoek een filosofische positie kan ondersteunen of verhelderen. De vraag is of dergelijke ondersteuning ook voor de ontwerphouding gevonden kan worden. Ook naar de ontwerphouding is recentelijk vrij veel onderzoek gedaan, overigens direct geïnspireerd door Dennett's werk erover. Hierbij richt het onderzoek zich vooral op technische objecten.

Dit onderzoek staat in de kinderschoenen maar de resultaten wijzen er vooralsnog op dat mensen in hun eerste tien levensjaren de ontwerphouding (*x*-is-voor-*y* typen redeneringen) ontwikkelen. Dit begint met het kunnen onderscheiden van interne en externe functies (ongeveer 3 jaar oud). In de jaren daarop beginnen kinderen technische objecten te classificeren in termen van oorspronkelijk bedoeld ontwerp: als iets ontworpen is voor doel *x*, dan is zijn functie doel *x* te vervullen en wordt het object als zodanig geclassificeerd. Denken in termen van oorspronkelijk bedoeld ontwerp veronderstelt een tamelijk goede beheersing van de intentionele houding en ontwikkelt zich rond een jaar of vier, vijf. Bijvoorbeeld, vanaf die leeftijd wordt een object waarvan verteld werd dat het ontworpen was als theepot, maar nu (al dan niet herhaaldelijk) gebruikt wordt om de planten mee water te geven, geclassificeerd als theepot. Het object ondersteunt beide functies even goed. Een stuk papier in de vorm van een hoed werd vaker geclassificeerd als een hoed als het als hoed gemaakt was, dan wanneer de vorm door toeval ontstond, omdat bijvoorbeeld een auto over het papier heen reed. Objecten die voor een bepaald doel gemaakt waren, maar nu stuk waren - bijvoorbeeld een kapotte vork - werden nog altijd als vorken gezien, ook al konden ze die functie niet meer vervullen.

Een of twee jaar later gebruiken kinderen deze theoretische kennis ook steeds meer in de praktijk. Ze zijn meer en meer geneigd om artefacten in termen van oorspronkelijk bedoeld ontwerp te classificeren. Dit wordt ook wel functie-fixatie genoemd. Dit betekent dat kinderen moeite krijgen om een artefact voor een a-typische (niet oorspronkelijk bedoelde) functie te gebruiken. Overigens passen kinderen *x*-is-voor-*y* redeneringen op alle mogelijke soorten objecten toe. Niet alleen technische ontwerpen, maar ook natuurverschijnselen (regen is om nat te maken), natuurlijke objecten (de steen is voor het dier om zijn kop aan te krabben, zand is korrelig zodat het niet wegwaait) en zelfs dieren (de leeuw in de dierentuin is om naar te kijken).

Niet alleen kinderen, ook volwassenen zijn sterk geneigd technische artefacten in termen van oorspronkelijk bedoeld ontwerp te duiden. Beslissend is dit onderzoek nog niet, maar de conclusie lijkt gerechtvaardigd dat mensen alleen van dit principe afwijken als objecten de oorspronkelijk bedachte functie echt niet kunnen of hadden kunnen vervullen. Dit gaat recht in tegen Dennett's claim dat de ontwerphouding niet in termen van intenties werkt. Ondersteuning voor de claim dat mensen in eerste instantie volgens optimaliteitsprincipes redeneren, is er nauwelijks (maar is ook niet expliciet onderzocht). Het empirisch onderzoek wijst vooralsnog in elk geval veel meer in de richting van intententionalisme, dan in de richting van optimaliteit.

Het ziet er, kortom, naar uit dat we in het alledaagse leven technische artefacten op een intentionele manier benaderen. De theorie is dan ook dat de ontwerphouding een verfijning en verdere ontwikkeling is van de intentionele houding. Mensen zijn intentionalisten die overal bedoelingen, intenties en betekenissen in en achter zien. De vraag waarom dat zo is, en waarom we ontwerpen in termen van intentioneel ontwerp zien, blijft voorlopig onbeantwoord in de empirische literatuur, maar er wordt wel gespeculeerd dat het te maken kan hebben met de snelheid waarmee beslissingen genomen kunnen worden. Dat is een interessante waarneming, die mooi aansluit bij de gedachte dat de intentionele houding ook vanwege haar vermogen snelle beslissingen te nemen, zo'n belangrijke rol inneemt in het menselijk leven.

Ik scheef al dat het empirisch onderzoek weinig ondersteuning geeft aan Dennett's optimaliteitsaccount (en dat terwijl het onderzoek direct door zijn theorie geïnspireerd is). Ook meer conceptueel gerichte filosofen moeten dit soort onderzoek ter harte nemen. Veel filosofische conceptualiseringen zijn gebaseerd op de manier waarop we in het alledaags leven termen gebruiken. Empirisch-wetenschappelijk onderzoek levert systematische en representatieve gegevens die naar mijn mening betrouwbaarder zijn dan het materiaal dat filosofen in hun leunstoel thuis kunnen verzamelen.

7. De ontwerphouding als methode

Dat de ontwerphouding uitgewekt als aan de hand van het optimaliteitsprincipe weinig empirische ondersteuning heeft, is slecht nieuws voor Dennett. Maar Dennett zou nog kunnen laten zien dat zijn account veel bruikbaarder is dan een intentionalistisch account. Voor Dennett zelf is dan op de eerste plaats wetenschappelijk-methodologische bruikbaarheid een belangrijke mogelijke waarde. Maar praktische bruikbaarheid zal ook geëvalueerd moeten worden.

De wetenschappelijk-methodologische bruikbaarheid van een liberale *intentionele* houding is door Dennett zelf herhaaldelijk aangetoond. Ze kan nog versterkt worden door te refereren aan bijvoorbeeld onderzoek in de cognitieve ethologie, de studie naar het denken en kennen van dieren. Hierin vervullen termen als 'gedachte', 'behoefte' en 'intentie' een waardevolle en bruikbare heuristische rol om iets te kunnen zeggen over hoe het dier zijn omgeving 'representeert' en welke 'doelen' het wil bereiken. Dergelijke termen worden dan vaak als functionele termen gezien (vgl. hoofdstuk 4), die generaliseren over een grote hoeveelheid instanties van gedrag die een bepaalde functie gemeenschappelijk hebben. Denken in termen van functies helpt ook bij het specificeren van 'inhouden' van mentale toestanden. Een mooi voorbeeld is dat van een zeker fruitvliegje, dat alleen kan overleven in een vochtige omgeving. Het fruitvliegje kan niet direct vochtige omgevingen detecteren, maar het heeft wel lichtsensoren, waardoor het op zoek kan gaan naar donkere plekken (die in zijn leefomgeving meestal ook vochtig zijn). De lichtsensor traceert op causaal niveau lichtintensiteit; op functioneel niveau zoekt het vochtigheid en 'denkt': 'há daar is vocht!'. Het gaat niet om de vraag of het vliegje 'echt' denkt – met alle discussies over bewustzijn van dien - maar om de rol die de term 'denken' kan spelen in het begrijpen van zijn gedrag.

In hoeverre kan de ontwerphouding als optimaliteitsaccount ook zo'n bruikbare rol spelen? Het is lastig aan te tonen dat het principieel niet kan - je kunt immers altijd een mogelijkheid over het hoofd zien - dus ligt het voor de hand een kritische case te nemen. Een voor de hand liggend wetenschappelijk gebied waarbij we geïnteresseerd zouden kunnen zijn in een methodologische toepassing van de ontwerphouding is de archeologie van de prehistorie, de artefactenstudie bij uitstek. Archeologie is een goede kritische case omdat er weinig gegevens zijn over de mentale capaciteiten van de eerste mensen op aarde, en de optimaliteitsaccount daar dus in het voordeel zou moeten zijn. En inderdaad lijkt het type redeneren van archeologen erg op het soort redeneringen dat de optimaliteitsaccount gebiedt: uitgaan van de best mogelijke rol die het object kan vervullen, en op basis daarvan iets zeggen over wat de eerste mensen ermee bedoelden. De welbekende vuistbijl en speculaties welke functie dit artefact gehad moet hebben, is hiervan een goed voorbeeld en lijkt de intentionele benadering tegen te spreken: naar alle waarschijnlijkheid zijn deze artefacten niet bewust gemaakt: onze voorouders waren wellicht zelfs niet in staat om dergelijke intenties te vormen. Dit noem ik het 'geen-blauwdruk argument'. Daarnaast blijkt dat functieredeneringen adequaat verlopen zonder referentie naar intenties. Een leuk voorbeeld is de eigenzinnige archeoloog Calvin, die tracht aan te tonen dat de vuistbijl helemaal geen *bijl* kon zijn, omdat de randen veel te scherp zijn om hem stevig vast te houden. De 'bijl' blijkt in experimentele situaties bovendien veel beter te werken als een soort discus om kuddedieren mee te verwonden. Een optimaliteitsredenering pur sang. Exit intentionalisme? Nee, want de argumenten zijn niet doorslaggevend.

Het geen-blauwdruk argument is niet voldoende om intentionalisme mee te verwerpen. Niet voor Dennett althans. Als onze voorouders niet in staat waren tot het vormen van complexe intenties, wil dat niet zeggen dat we hun gedrag niet in termen van intenties kunnen duiden. Als we immers intentionele toestanden aan thermostaten mogen toeschrijven, dan toch zeker ook aan deze eerste mensen, die cognitief zeker tot meer in staat waren dan een simpele thermostaat!

Het argument dat archeologische redeneringen (à la Calvin) goed toe kunnen zonder referentie aan intenties, is evenmin beslissend. Degelijker redeneringen stoelen namelijk op een brede set van veronderstellingen over de intentionele context. Als we zouden weten, bijvoorbeeld, dat de vroege mensen geen vlees wilden eten, zou dat de discus-hypothese zwaar onder druk zetten.

Beslissend is deze argumentatie niet, maar het lijkt er toch sterk op dat functionele verklaringen (de ontwerphouding) sterker zijn als we informatie hebben over de intentionele context waarin het object gemaakt en gebruikt hebben. Een prachtig voorbeeld hierbij is de beroemde Babylonische Batterij, een merkwaardig object dat uitstekend stroom blijkt te kunnen opwekken (wat voor sommigen reden is om te veronderstellen dat er buitenaards leven op aarde moet zijn geweest). Stroomopwekking is zeer waarschijnlijk niet de functie waarvoor de Babylonieërs het ding gebruikten. Meer waarschijnlijk past het in de religieuze overtuigingen van de Babylonieërs en was het bedoeld om heilige geschriften veilig in op te bergen: een functie die alleen achterhaald kan worden door iets te weten over het mentale (religieuze) leven van de Babylonieërs. Zo zijn er vele voorbeelden van mogelijke (optimale) functies van archeologische objecten, die verworpen moeten worden op basis van de intentionele context. Dennett erkent dit soms, maar als hij dit echt vindt, zou hij zijn optimaliteitsprincipe moeten laten vallen.

Er zijn nog twee argumenten tegen de optimaliteitsbenadering van de ontwerphouding. Dennett stelt dat de intentionalistische benadering slecht voorspelt. Een object kan immers een andere functie krijgen. En de historische functie, stelt Dennett, is niet interessant. Merkwaardig is deze opvatting wel, zeker omdat Dennett zelf graag voorbeelden gebruikt uit het verleden. Maar deze tegenstrijdigheid even daargelaten, is het geen sterk argument voor de optimaliteitsbenadering. Optimaliteit is een contextuele notie, een object is optimaal voor het bereiken van een bepaald doel gegeven de huidige situatie. Als zodanig valt er met optimaliteit bar weinig te voorspellen. Zowel intentionalistische als optimaliteitsbenaderingen van de ontwerphouding zijn slechte voorspellers, maar intentionalistische benaderingen zijn leveren sterkere uitspraken over de historische functie van een object. De intentionalistische benadering lijkt dus ook wat dit betreft nuttiger dan de optimaliteitsbenadering.

Misschien is het niet de bedoeling om de ontwerphouding voor wetenschappelijke doeleinden te gebruiken. Vragen over doelen horen wellicht niet in de wetenschap thuis. Zijn er dan andere mogelijke methodologische toepassingen van de ontwerphouding waarin de optimaliteitsbenadering zijn vruchten kan afwerpen? Ja, we zouden bijvoorbeeld kunnen kijken naar de methodologische waarde van de ontwerphouding voor alledaagse conceptualiseringen ten behoeve van moraal en recht. In een paper werkt Dennett bijvoorbeeld een sterke conceptualisering uit van het begrip 'persoon', waarbij hij zich laat leiden door praktische bruikbaarheid van het concept. Het begrip 'persoon' is een beladen begrip, dat belangrijke keuzes beïnvloedt. Opvattingen over abortus bijvoorbeeld, of het recht zelf te beslissen. Misschien helpt Dennett's notie van functie om bepaalde morele of juridische beslissingen te rechtvaardigen. Maar hier vermoed ik dat een intentionele benadering veel sterker en bruikbaarder is dan een optimaliteitsbenadering. De notie van intentie is fundamenteel voor moraal en recht, het helpt verschil maken tussen moord en doodslag, tussen onwaarheid vertellen en liegen, tussen een ongeluk en opzet. Dit zijn, in het alledaagse leven, fundamentele verschillen, die ook van belang zijn voor functietoeschrijving aan technische objecten. Denk alleen al aan het recht je dvd-speler te mogen retourneren als hij niet functioneert als dvd-speler. Dat de dvd-speler nog altijd als een uitstekende *presse papier* kan dienen, interesseert de normale consument geen fluit.

8. Conclusie

Intenties zijn, zeker vanuit objectief-wetenschappelijk perspectief, zeer problematische fenomenen. Waarschijnlijk zullen we nooit exact aan kunnen geven wat een intentie is. Hetzelfde geldt voor gedachten, wensen, en alle andere attitudes die de revue zijn gepasseerd. Maar het menselijk leven kan niet zonder. Mensen zijn intentionalisten en vaak is dat erg nuttig. Dat is misschien wel de essentiële gedachte achter Dennett's ideeën over de intentionele houding, maar opvallend genoeg laat hij deze los zodra hij over ontwerpen spreekt. Begrijpelijk is het wel, gezien Dennett's opvattingen over de ontworpen aard van de menselijke geest, en de dreigende circulariteit als ontwerpen in intentionele termen geduid zouden worden.

Helaas wringt Dennett's oplossing om de dreigende circulariteit te vermijden, het optimaliteitsprincipe, evenzeer met de rest van zijn theorie. De optimaliteitsprincipe blijkt nauwelijks in overeenstemming met de empirische realiteit, en de methodologische waarde is zwak, zeker in vergelijking met een intentionele benadering.

Het is de vraag of er een uitweg is uit deze klemmende situatie. Als mijn redeneringen kloppen, zal het moeilijk zijn de ontwerphouding kloppend te maken met de rest van de theorie. Want als Dennett zijn optimaliteitsprincipe vervangt voor een meer intentionalistische benadering, dan komt hij in de knel met theses 3 en vooral 2 zoals besproken in hoofdstuk 5.

Een intentionele benadering wringt met these 3, het principe van functiegeneraliteit, omdat dan ook biologische ontwerpen op intentionele wijze geïnterpreteerd zouden moeten worden. Op zichzelf is dit niet zo'n probleem, Dennett's liberale attitude theorie kan best (quasi-)intentie toeschrijvingen aan de natuur accommoderen. De echte pijn zit hem in these 2: primaat van (biologisch) ontwerp. Omdat Dennett intentionele toestanden wil verklaren met behulp van biologische ontwerpprincipes, kan hij biologische ontwerpen (en gezien these 2 dus ook technische ontwerpen) niet zomaar verklaren door naar intentionele toestanden terug te verwijzen. De uitdaging is om deze circulariteit op te lossen. Een filosofisch hoogstandje.

Curriculum Vitae

Melissa van Amerongen graduated cum laude at the University of Amsterdam in the Sociology of Knowledge and Science, a combined program in which the studies Philosophy and Sociology are combined. She graduated on the topic of lying (i.e. intentionally not telling the truth) with the thesis *Je loog tegen mij; een voorstudie naar de sociologie van het liegen* (2000). From 2001-2006 she worked at the Technical University of Eindhoven, as a Ph.D. student on her dissertation, *The Interpretation of Artifacts; A Critique of Dennett's Design Stance*. From 2005 she works as the assistant editor of the philosophical journal *Philosophical Explorations*. Since 2007, she also works as a researcher at the Stichting Kennisnet (Zoetermeer), where she is mainly involved in scientific research on effective use of ict in education.

Simon Stevin Series in the Philosophy of Technology
Delft University of Technology & Eindhoven University of Technology
Editors: Peter Kroes and Anthonie Meijers

Books and Dissertations

Volume 1: Marcel Scheele, '*The Proper Use of Artefacts: A philosophical theory of the social constitution of artefact functions*', 2005
Volume 2: Anke van Gorp, *Ethical issues in engineering design, Safety and sustainability*, 2005
Volume 3: Vincent Wiegel, SophoLab, *Experimental Computational Philosophy*
Volume 4: Jeroen de Ridder, *Technical Artifacts: Design and Explanation*, 2006
Volume 5: Melissa van Amerongen, The Interpretation of artifacts; A critique of Dennett's design stance, 2008

Research Documents

Peter Kroes and Anthonie Meijers (eds.), *'Philosophy of Technical Artifacts'*, 2005

Simon Stevin (1548-1620)

'Wonder en is gheen Wonder'

This series in the philosophy of technology is named after the Dutch / Flemish natural philosopher, scientist and engineer Simon Stevin. He was an extraordinary versatile person. He published, among other things, on arithmetic, accounting, geometry, mechanics, hydrostatics, astronomy, theory of measurement, civil engineering, the theory of music, and civil citizenship. He wrote the very first treatise on logic in Dutch, which he considered to be a superior language for scientific purposes. The relation between theory and practice is a main topic in his work. In addition to his theoretical publications, he held a large number of patents, and was actively involved as an engineer in the building of windmills, harbours, and fortifications for the Dutch prince Maurits. He is famous for having constructed large sailing carriages.

Little is known about his personal life. He was probably born in 1548 in Bruges (Flanders) and went to Leiden in 1581, where he took up his studies at the university two years later. His work was published between 1581 and 1617. He was an early defender of the Copernican worldview, which did not make him popular in religious circles. He died in 1620, but the exact date and the place of his burial are unknown. Philosophically he was a pragmatic rationalist for whom every phenomenon, however mysterious, ultimately had a scientific explanation. Hence his dictum 'Wonder is no Wonder', which he used on the cover of several of his own books.

According to the philosopher Daniel Dennett, the function of an artifact is what it is best able to do, regardless of what its designer intended it to be, or what users intend to use it for. Dennett's so-called optimality account of function is his alternative to intentionalistic and causal accounts. It is a crucial component of his theory. It allows Dennett to explain intentionality in terms of the notion of design and this justifies in turn his ideas about the intentional stance as being dependent upon the design stance.

This books investigates the consistency of the optimality account in relation to the rest of Dennett's theory. It analyzes whether the optimality account satisfies standards internal to his theory. The main thesis of the book is that the optimality account does not live up to these standards, thus undermining the very foundations on which it is built.

'Wonder en is
gheen wonder'

TU/e Technische Universiteit Eindhoven University of Technology

TU Delft
Delft University of Technology