Emotions as functional kinds:

A meta-theoretical approach to constructing scientific theories of emotions

Dissertation
zur Erlangung des akamemischen Grades

Doctor philosophiae
(Dr. phil.)

eingereicht
an der Philosophischen Fakultät
der Humboldt-Universität zu Berlin

von
Juan Raúl Loaiza Arias

Die Präsidentin der Humboldt-Universität zu Berlin: Prof. Dr.-Ing. Dr. Sabine Kunst
Die Dekanin der Philosophischen Fakultät: Prof. Dr. Gabriele Metzler

Gutachter
Erstgutachter: Prof. Dr. Michael Pauen (Humboldt-Universität zu Berlin)
Zweitgutachter: Prof. Dr. Jesse Prinz (City University of New York Graduate Center)
Drittgutacther: Prof. Dr. med. Dr. phil. Henrik Walter (Charité - Universitätsmedizin)

eingereicht am: 24.09.2019
verteidigt am: 13.07.2020

Abstract

In this dissertation, I address the question of how to construct scientific theories of emotions that are both conceptually sound and empirically fruitful. To do this, I offer an analysis of the main challenges scientific theories of emotions face, and I propose a meta-theoretical framework to construct scientific concepts of emotions as explications of folk emotion concepts.

Part I discusses the main challenges theories of emotions in psychology and neuroscience encounter. The first states that a proper scientific theory of emotions must explain all and only the phenomena under the vernacular term 'emotion' with a common set of conceptual resources and under an overarching generic concept of emotion. The second demands that each emotion category corresponds to well-coordinated sets of neural, physiological, and behavioral patterns of responses. I argue that none of the best contemporary theories of emotions in psychology and neuroscience overcomes these challenges. As a result, a new theory of emotions is required.

In Part II, I develop the meta-theoretical framework to construct a theory of emotions that overcomes the challenges above. First, I propose a pluralistic account of scientific kinds based on different patterns of projection that various disciplines may take to justify inductive inferences. These are essentialist, historical, and social patterns. Each of these patterns provides a framework to construct different types of scientific concepts. Second, I argue that to decide between the different frameworks of scientific kinds to construct tractable theories of emotions, we must engage in what Bechtel and Richardson call "reconstituting the phenomena." I suggest that in the case of emotions, reconstitution amounts to the explication of folk emotion terms in the vocabulary specified by the target framework. Lastly, I argue that among the frameworks for scientific kinds available, the one that is best suited to explicate emotion concepts is a functional framework. Consequently, I conclude by recommending scientists pursue functionalist theories of emotions over essentialist, historical, or social theories.

Zusammenfassung

In dieser Dissertation beschäftige ich mich mit der Frage, welchen Anforderungen wissenschaftliche Theorien über Emotionen gerecht werden müssen, damit sie sowohl begrifflich fundiert als auch empirisch fruchtbar sind. Zu diesem Zweck biete ich zunächst eine Analyse der wichtigsten Herausforderungen, mit denen wissenschaftliche Emotionstheorien konfrontiert sind. Anschließend schlage ich einen metatheoretischen Rahmen vor, in dem wissenschaftliche Konzepte von Emotionen als Begriffsexplikationen von Alltagsemotionskonzepten konstruiert werden können.

Teil I diskutiert die wichtigsten Herausforderungen für Theorien der Emotionen in der Psychologie und den Neurowissenschaften. Die erste Herausforderung ist, dass eine wissenschaftliche Theorie der Emotionen alle und nur die Phänomene unter den Alltagsbegriff „Emotion“ subsumieren sollte, die durch gemeinsame begriffliche Ressourcen erfasst werden können. Die zweite Herausforderung ist, dass jede Emotionskategorie gut koordinierten Gruppen neuronaler, physiologischer und verhaltensbezogener Reaktionsmuster entsprechen sollte. Ich behaupte, dass keine der derzeitigen Theorien der Emotion in Psychologie und Neurowissenschaft dieser Anforderung entspricht. Infolgedessen ist eine neue Theorie der Emotionen erforderlich.

Teil II entwickelt den metatheoretisches Bezugssystem für eine Theorie der Emotionen, die den oben genannten Herausforderungen entspricht. Erstens schlage ich eine pluralistische Darstellung der Kategorien oder „scientific kinds“ vor, die induktive Schlussfolgerungen begründen können. Jedes dieser Muster bietet einen Rahmen, um verschiedene Arten von wissenschaftlichen Konzepten zu konstruieren. Zweitens behaupte ich, dass wir uns, um zwischen den verschiedenen wissenschaftlichen Bezugssystemen wissenschaftlicher Kategorien zu entscheiden, mit dem beschäftigen müssen, was Bechtel und Richardson „reconstituting the phenomena“ nennen. Im Falle von Emotionen muss die Rekonstitution in der Explikation von Alltagsbegriffen der Emotion in demjenigen Vokabular erfolgen, welches das relevante Bezugssystem bereitstellt. Abschließend argumentiere ich, dass das funktionale Bezugssystem für wissenschaftliche Kategorien oder „scientific kinds“ am besten zur Erläuterung von Emotionskonzepten geeignet ist. Folglich schließe ich mit der Empfehlung, dass Wissenschaftler*Innen funktionalistische Theorien von Emotionen anstelle von essentialistischen, historischen oder sozialen Theorien in ihren Studien benutzen sollten.

Contents

Abstract iii

Zusammenfassung v

Contents vii

List of Figures xi

List of Tables xiii

Acknowledgments xv

Introduction 1

I The Problem 5

1 The Theoretical Challenge: The Problem of Disunity 7
- 1.1 The Theoretical Challenge: Griffiths's eliminativism . . . 8
- 1.2 Updating the Theoretical Challenge . . . 11
  - 1.2.1 Basic Emotion Theory . . . 12
  - 1.2.2 Appraisal Theory . . . 28
  - 1.2.3 Psychological Constructionism . . . 35
- 1.3 Prospects to overcome the Theoretical Challenge . . . 43

2 The Empirical Challenge: The Problem of Variability 45
- 2.1 The Variability Thesis . . . 46
  - 2.1.1 Clarifying VT . . . 48
- 2.2 Individuating response patterns . . . 54
  - 2.2.1 Neural patterns . . . 54
  - 2.2.2 Physiological patterns . . . 69
  - 2.2.3 Behavioral patterns . . . 74
  - 2.2.4 Expressive patterns . . . 76
  - 2.2.5 Phenomenological patterns . . . 83
- 2.3 How to understand variability? . . . 86

3 Interlude: Can we study emotions scientifically? 91
3.1 The Two Challenges Considered . . . 92
3.1.1 Where the challenges meet . . . 92
3.1.2 Where the challenges diverge . . . 93
3.2 Failing to meet the challenge: Should we just give up? . . . 95
3.3 What do we need to theorize about emotions? . . . 97

II The Solution 99

4 Scientific Kinds 101
4.1 The tradition of natural kinds . . . 102
4.1.1 Mill and the introduction of Kinds . . . 102
4.1.2 Essentialism . . . 106
4.1.3 The HPC account . . . 108
4.2 Projectibility and scientific kinds . . . 117
4.2.1 Realism and kinds . . . 120
4.2.2 Giving up 'natural' . . . 123
4.3 A taxonomy of scientific kinds . . . 124
4.3.1 Essential kinds . . . 125
4.3.2 Historical kinds . . . 126
4.3.3 Functional kinds . . . 127
4.3.4 Social kinds . . . 130
4.4 Conclusion . . . 133

5 Taking a step back: Reconstitution through explication 135
5.1 What is reconstitution? . . . 136
5.1.1 Reconstitution and mechanistic explanations . . . 136
5.1.2 Bechtel and Richardson on Mendel . . . 138
5.2 Examining reconstitution . . . 142
5.3 How to carry out reconstitution . . . 144
5.4 Explication . . . 146
5.4.1 Strawson contra Carnap . . . 148
5.4.2 Scarantino's two emotion projects . . . 150
5.5 Conclusion: How to explicate emotions . . . 155

6 A Functionalist Approach to Emotion Kinds 159
6.1 The Folk-Psychology of Emotions . . . 159
6.1.1 (Relatively) Uncontroversial features . . . 160
6.1.2 Context-sensitivity . . . 162
6.1.3 Minimal attribution and non-human animals . . . 166
6.2 How do emotions fit together? . . . 169
6.2.1 Considering the relevance of folk-psychological features . . . 170
6.2.2 What emotions are not . . . 172

6.2.3 Why a functional model? 182
6.3 Objections to functionalism 187
6.3.1 Functionalism is not falsifiable 187
6.3.2 Functionalism is teleological 190
6.4 Conclusion 194

7 Concluding remarks **197**
7.1 Responding to the challenges 199
7.1.1 Revisiting the Theoretical Challenge 199
7.1.2 Revisiting the Empirical Challenge 202
7.2 Future avenues for a science of emotions 203

References **205**

Statement of Authorship **219**

List of Figures

2.1 Variations of NOC 47

5.1 Scarantino's view of the relation between folk emotion terms and natural kinds of emotions. 153
5.2 Sketch of the difference between folk and scientific emotion concepts. . 156

List of Tables

1.1 Roseman's emotion families. 31

2.1 Summary of results of the discriminability analysis by Vytal and Hamann (2010). 58

Acknowledgments

As in every other project, there were many who contributed either directly or indirectly. Of those who contributed directly, I would like to begin by thanking my supervisors. First, thanks to Michael Pauen, who acted as the main supervisor of this work and whose strict but helpful feedback shaped many of the ideas presented here. Second, to Jesse Prinz, who acted as co-supervisor and whose enthusiasm and encouraging feedback were paramount to completing this dissertation. There is no philosopher I have met that is so passionate and who enjoys discussing as much as Jesse. Lastly, thanks to Isabel Dziobek, who warmly invited me into her Social Cognition Lab (now Clinical Psychology of Social Interaction Group), where I learned many lessons about actual scientific practice to which I have tried to do justice here.

Second, I woud like to express my gratitude towards my colleagues and friends at the Berlin School of Mind and Brain. To the Social Cognition Lab researchers, who not only taught and showed me the ups and down of scientific practice, but also welcomed me and invited me to discuss my work and ideas with them and allowed me to contribute to theirs. It is the dream of any interdisciplinary researcher to have access and support from such a group. Among them, I would like to mention especially and in no particular order: Caitlin Duncan, Lena Matyjek, Jan Schneider, Garret O'Connell, Simón Guendelman, Mareike Bayer, and Hanna Drimalla. I would also like to thank other colleages at Mind and Brain as well as fellow visiting academics whose feedback also helped polish this work. These are Dimitri Coelho-Mollo, Matteo Colombo, Javier Gómez-Lavin, Astrid Schomäcker, Gen Eickers, Robin Guido Löhr (who also helped translate the abstract), and Esra Al, as well as other members of the Philosophy colloquium. Also, thanks to Johannes Mohn for his additional help with the translation of the abstract.

Third, a very special word of my most profound gratitude to Diana Pérez at the Instituto de Investigaciones Filosóficas of the Sociedad Argentina de Análisis Filosófico (SADAF) in Buenos Aires, Argentina. Diana received me as a visiting researcher in Buenos Aires during the last stages of this dissertation. During this time, Diana discussed and challenged every argument presented in this work, both with empathy and tenacity. This was both an inspiring experience as well as an example of how academia should be. Additionally, I would like to mention members of Diana's group who also welcomed me in Buenos Aires and with whom I expect to keep collaborating in the future. These are Andrea Melamed, Lucas Bucci, Federico Burdman, Karina Pedace, and Tomas Balmaceda.

Fourth, I would like to thank those whose emotional support made it possible to start and complete this work. In the first place, I would like to thank Irene Trilla, who not only contributed with her sharp scientist mind to the ideas developed here, but also helped me stand up when times were hard and lived many of the emotions involved in this project alongside with me. Second, I want to thank my family, who always supported my project of coming to Berlin to pursue my master's and now my doctorate. These are my parents, Francia Helena Arias and Adalberto Loaiza, let this work be a testament of their effort; and my siblings, Laura, Lara Susana, and Marcos.

In addition to those above, there are many others whose contribution doesn't fit into any of the other categories but whose minds had an impact on this project. These are, first, my emotion researcher colleagues and friends, Marco Viola (Università degli Studi di Torino) and Rodrigo Díaz (Universität Bern), with whom I was lucky to share a passion for these problems regarding a science of emotions. I hope that this is just the start of many collaborations and philosophical discussions with them. Second, to my former mentor, Carlos Cardona at the Universidad del Rosario in Bogotá, Colombia, who gave me the most important piece of advice I could have received for my doctoral studies: "Cuando haya dificultades (sin duda las habrá) compense el desasosiego que ello produce con altas dosis de disciplina" (when difficulties arise (they will without a doubt), make up for the anxiety that that produces with high doses of discipline). Third, I would like to thank Henrik Walter for accepting to write the third review of this dissertation and for his input and questions on the review and during the defense. Lastly, thanks to all other friends that might be missing here.

Introduction

Emotions have been an object of both scientific investigation and philosophical thought for centuries. The first attempts to offer a scientific theory of emotions can be traced back to the efforts of Wundt (1897) and, perhaps most famously, James (1884). In turn, philosophical thinking about the emotions can be found, *pace* the possible anachronisms, since Aristotle (n.d./2009, n.d./2018), going through Descartes (1649/1985), Spinoza (1677/2018), and Hume (1738/2000), to mention a couple.

Even though these two histories (the philosophical and scientific ones) might have occurred in a somewhat independent manner, recent decades have seen their integration. Specifically, philosophers have become increasingly aware of the need to inform their thinking by empirical findings. Conversely, scientists have become aware of the need of conceptual clarification if they are to be successful at understanding what emotions are. This work situates itself in this interdisciplinary area, connecting insights from psychology and neuroscience into a philosophical framework that, in turn, expects to contribute to scientific enterprises.

The exact problem this work intends to contribute to can be summarized—albeit too generally—under the question posed by Scarantino (2012): how can we define emotions scientifically? In the past years, a growing amount of philosophical thinking and empirical literature has suggested that a scientific definition of emotions, at least as traditionally attempted, is a project that is bound to fail. The main reason behind this pessimism is that emotions, it seems, do not constitute natural kinds. If this is so, pessimists argue, emotions cannot be legitimate objects of scientific study, or if they can, it cannot be under a common theoretical framework that explains all emotions and only the emotions. In any case, formulating a unified, scientific theory of emotions seems implausible.

As I explained above, the sources of this pessimism come both from philosophy as well as from scientific literature. In philosophy, the most influential defense of this pessimism can be found in Griffiths (1997). According to Griffiths, none of our best theories of emotions captures all of the phenomena we call emotions in our everyday lives. Instead, these theories suggest that the overarching concept of EMOTION includes different kinds of phenomena, each explained by its own theory. If this is true, efforts to offer a unified theory of emotions are futile. Emotions, Griffiths argues, are just too heterogeneous to fit into a common theoretical framework. The best we can do is work in different theories, each explaining only a subset of the phenomena captured by the overarching concept.

On the side of empirical literature, the most ardent defense of pessimism comes from psychological constructionists, particularly in the work of Lisa Feldman Barrett (2006; 2017; 2018a). For constructionists such as Barrett, empirical research has already found that emotions do not form a natural kind. Evidence from neuroscience and psychology shows, according to them, that emotion categories are too heterogeneous, and that rather than finding unity, "variation is the norm" (Barrett, 2018a, p. 23). As a result, they propose a theory of emotions which does not demarcate between emotions and other affective phenomena, and that makes distinctions between emotion categories a matter of concepts rather than facts.

These two lines of arguments for pessimism raise important issues for emotion research. First, they call into question the possibility of studying emotions scientifically. Second, they invite a philosophical but empirically informed investigation into what is it that scientists are after when investigating emotions. This dissertation takes an important step towards offering answers to these problems. It intends to resist these forms of pessimism and argue that we can formulate a scientifically meaningful theory of emotions.

To do so, this work makes two important contributions to the existing debate on emotions and their scientific study. On the one hand, this work provides an analysis of the challenges scientists face when proposing scientifically meaningful theories of emotions. On the other hand, and with this analysis in hand, this work proposes a strategy to overcome the aforementioned challenges. Specifically, here I propose a meta-theoretical framework from which scientists can formulate what I think can be successful scientific theories of emotions. This framework invites scientists to construct scientific emotion concepts as explications of folk emotion concepts in functional terms. While I shall not fill in the blanks and propose a specific theory of emotions here, the desiderata I present as part of this framework are intended to direct science in a direction that avoids the pitfalls of previous theories.

The dissertation is divided into two parts. Part I focuses on the challenges mentioned above, the challenges scientists face when proposing scientific theories of emotions. These are what I call the *Theoretical Challenge* and the *Empirical Challenge*.

The Theoretical Challenge stems from Griffiths's arguments mentioned above. It puts constraints on the scope and extension of scientific theories of emotions, claiming that a satisfactory theory must provide a systematic theoretical framework that is empirically tractable and that includes all and only the phenomena under the vernacular term "emotion." Chapter 1 is devoted to an update and examination of this challenge from the perspective of contemporary emotion research in psychology and neuroscience. It discusses the best theories of emotions to this day, and argues that as the literature stands, no theory yet has succeeded in meeting the challenge.

The Empirical Challenge, in turn, stems from constructionist arguments and empirical findings. Specifically, I trace this challenge back to what Scarantino (2015) calls the *Problem of Variability*, and which I call the *Variability Thesis*. The Variability Thesis claims that emotions are naturally disjoined phenomena. It is supported by empirical findings showing that there are no coordinated patterns of responses corre-

sponding to emotions, and if there are, that there are no one-to-one correspondences between these and emotion categories. In chapter 2, I examine this thesis in detail. First, I argue that as it stands, this thesis is problematically ambiguous, and that further clarification is required in order to understand its empirical import and thus for it to be useful in the interpretation of empirical evidence. After I propose an analysis of this thesis, I examine empirical findings in its support. I conclude that while there are good reasons to accept this thesis, much of its empirical import still depends on our theoretical commitments regarding how to individuate neural, physiological, and behavioral patterns.[1]

In chapter 3, I discuss the two challenges together. In this interlude, I examine some points of convergence and divergence between these challenges, and examine their presumed consequences. In particular, this chapter raises the question of what follows from past failures to meet these challenges. I reject the eliminativist conclusion that emotions should be taken out of the set of legitimate objects of scientific investigation, and suggest a revision of some of our meta-theoretical assumptions regarding how we think about emotions in the scientific domain. This revisionist suggestion sets the agenda for the second part of the dissertation.

Having analyzed the problem in detail, Part II sets out to propose the aforementioned meta-theoretical framework to formulate a scientific theory of emotions. Broadly construed, the claim I will defend is that a viable strategy to meet the aforementioned challenges is to explicate folk emotion kinds into scientifically meaningful functional kinds. This makes it possible to overcome the Theoretical Challenge by enabling the formulation of a satisfactory theory of emotions while dissolving the Empirical Challenge as traditionally understood by allowing the postulation of multiply realizable but empirically interesting kinds.

To develop this strategy, Part II is divided into three main steps. First, I suggest that part of the problem leading to the challenges above is an overly restrictive account of natural kinds. In chapter 4, I examine the history of the notion of natural kinds leading up to this restrictive account, and argue that what is important about natural kinds in science is how they answer questions about how to justify inductive inferences and projectibility. I distinguish four types of what I call *scientific kinds*, each corresponding to different ways in which scientific disciplines justify inductive inferences and projectibility. These are essentialist, historical, functional, and social kinds. With this pluralistic account of kinds in place, I claim that questions about whether emotions form natural kinds or not should be reframed in terms of what type of model of scientific kinds allows us to capture the phenomena more accurately while allowing the formulation of projectible categories. In other words, the question becomes what type of kinds do emotions form.

To answer this question, I suggest that we must take a step back and recharacterize the phenomenon that a scientific theory of emotions must explain. Put differently, we must recharacterize our explanandum. To do this, I submit, we must engage in what has been called in the literature on mechanistic explanations *reconstituting the*

[1] Parts of this chapter were published in Loaiza (2020).

phenomena. What is reconstitution and how it can apply to the case of emotions is the subject of chapter 5. I propose an account of reconstitution consisting of two steps. In the first step, scientists must find concepts to make reference to the phenomenon of interest. These concepts, which may be pretheoretical, play an ostensive role, allowing scientists to agree on the explanandum phenomenon. Once these concepts are in play, the second step is to find a framework to construe the explanandum as a scientific kind. In the case of emotions, I submit that the concepts that aid in fixing the explanandum are folk emotion concepts, and the task of finding a framework to construe them in terms of scientific kinds takes the form of *explicating* folk emotion concepts.

Lastly, in chapter 6, I apply the strategy above to evaluate which framework or model of scientific kinds best fits our pretheoretical concepts of emotions. I argue against essentialist, historical, and social models, and defend that emotions are best understood as functional kinds. Put differently, I claim that emotions are best explicated in terms of relations between inputs, outputs, and other mental states, rather than in terms of an essential property, a common causal history, or as intrinsically social practices. This way of cashing out emotion kinds functionally has clear advantages over other models, including allowing for multiple realization while preserving the unity of emotion categories. As a result, I submit, functional models can help scientists overcome the Theoretical Challenge by offering a framework that preserves conceptual and explanatory unity. Simultaneously, functional models allow for a dissolution or a reinterpretation of the Empirical Challenge, since theories following this framework can integrate variability without running into the issues previous theories of emotions faced. If this approach is correct, then the best way for scientists to define emotions scientifically is by adopting a functional approach.

Before closing this introduction, let us go over some clarificatory notes about the scope and target of this dissertation. As I explained above, this work does not offer a specific theory of emotions, nor does it defend a pre-existing theory. All of the claims presented and defended here are intended as meta-theoretical claims, that is, as claims about how to propose theories in general. While I will discuss specific theories in various points of this work, I shall only do so to point out past problems or to clarify why the framework I am proposing offer a more promising route for scientific research on emotions. Only in closing will I consider the prospects of current theories in the light of the framework I will propose.

It is also worth pointing out that this work will not discuss philosophical theories of emotions such as those defended by Solomon (2003), Nussbaum (2001), and others. This is because I am interested in theories of emotions in psychology and neuroscience specifically. While some philosophical theories might have bearing on scientific work (see e.g. Prinz, 2004), I will treat philosophical work on emotions as a separate subject matter from scientific theorizing. It remains an open question to which extent the framework developed here has an impact on these theories or whether these theories have a role to play with regards to work in psychology and neuroscience.

Part I

The Problem

Chapter 1

The Theoretical Challenge

The Problem of Disunity

There is a huge variety of theories of emotions. Different theories naturally raise different expectations. Some theories take emotions to be hardwired circuits ingrained in our biology, whereas for others claim they are constructions of our psychological makeup or our social environment, with little to nothing biological to them. Some theories expect emotions to be neatly organized into a finite set of categories, while others submit that there are potentially infinite possible ways to categorize them. There are even theories that reject the view that emotions form an interesting scientific kind in the first place, while others hold on to this view and stress the epistemic value of emotion categories.

In the introduction of this work, I presented what I called Griffiths's challenge, the challenge of finding a theory of emotions that presents emotions as scientific kinds. In this chapter, I will examine this challenge in detail. I will explore the question of whether any of the main current theories of emotions meets the challenge, updating Griffiths's arguments and evaluating their status in today's landscape. To do this, we must first go over the main theories on the market and evaluate whether they implicitly or explicitly present emotions in a scientifically interesting way. I will claim that none of the current theories meet the challenge completely, but some of them do offer important clues in that direction. Hence, even though we need a new theory of emotion kinds, there is still hope in this project.

Evaluating the different theories of emotions involves a number of steps. First, I present Griffiths's original argument and introduce what I call *the Theoretical Challenge*. Broadly construed, the theoretical challenge consists in offering a theoretical framework that explains all emotions and only the phenomena we call emotions with the same conceptual framework. In its original formulation, the theoretical argument considers only a subset of current emotion theories. Hence, in the second section, I update the challenge by considering contemporary empirical theories of emotion in psychology and neuroscience. These are basic emotion theories, appraisal theories, and psychological constructionism. I will explain the main tenets of each one of these views, along with their most salient variations. I then evaluate whether these theories

meet the theoretical challenge, arguing that none of them do the job. As a result, a new theory of emotions is required if we are to study emotions scientifically.

1.1 The Theoretical Challenge: Griffiths's eliminativism

Debates on the natural kind status of emotion date back to Griffiths's (1997) influential work. In his book, Griffiths claims:

> My central conclusion is that the general concept of emotion is unlikely to be a useful concept in psychological theory. It is meant to be a kind of psychological process that underlies a certain range of human behaviors. But there is no one kind of process that underlies enough of this behavior to be identified with emotion. (Griffiths, 1997, p. 14)

In Griffiths's view, the concept EMOTION is comparable to other concepts in science which have proved unfruitful, such as SUPRALUNARY OBJECT. Concepts of this sort have nothing scientifically meaningful in common, thus precluding them from figuring in interesting inferences and research agendas. Just as with other concepts of this type, Griffiths concludes that we must eliminate emotion from scientific projects altogether.

According to Griffiths, a successful theory of emotions would be one that refines or replaces emotion concepts "so that the categories corresponding to emotion concepts have strong causal homeostasis" (Griffiths, 1997, p. 228). By identifying mechanisms subserving causal homeostasis (the co-occurrence of observable properties; see Boyd, 1999a), science would be able to account for what grounds inferences across emotion categories. For instance, if the best theory of emotions identified them with activity in specific brain regions, this would enable us to infer properties from observed instances of a given emotion to unobserved instances in virtue of this common mechanism. This would provide the foundations of a scientifically tractable theory of emotions. Nonetheless, Griffiths argues, such project is bound to fail.

Griffiths's reasons to be pessimistic about the scientific status of emotions lie in the heterogeneity of the phenomena that fall under this category. In our folk-psychological, pretheoretical vocabulary, EMOTION is a category that ranges from instinctive responses like fearing a snake to complex ones such as feeling ashamed that we have failed to meet a moral norm. As mentioned above, any successful theory of emotions must explain all of these phenomena under the same framework. Yet, Griffiths argues, none of our best theories can do so. In all cases, some subset of phenomena are left unaccounted for.

Griffiths identifies two natural classes of emotions. One is the range of phenomena explained in terms of affect programs. As Griffiths understands them, affect programs are emotional responses that exhibit three properties: complexity, coordination, and automaticity. Complexity means that there are different elements constituting these responses, such as facial expressions, musculoskeletal responses, vocal changes, hormonal changes, and autonomic activity. Coordination means that these different elements occur together in specific patterns. Lastly, automaticity means that this coordination

occurs involuntarily and without conscious control. Furthermore, affect programs also exhibit a modular structure, i.e. they are informationally encapsulated in order to provide quick responses and to support their automaticity. Canonical examples of emotions of this type include instances of surprise, fear, anger, disgust, sadness, and joy, among others. In the next section, I present this view in detail.

The second type of phenomena under the concept of emotion is what Griffiths calls the "higher cognitive emotions." Unlike the emotions explained as affect programs, higher cognitive emotions demand a more complex cognitive architecture that escapes the conceptual resources of the first class. As Griffiths puts it, higher cognitive emotions do not involve the "brief, highly stereotyped emotional reactions" (Griffiths, 1997, p. 100) that affect programs do. Additionally, these emotions are neither informationally encapsulated, nor are the actions that unfold automatic and involuntary. Nonetheless, they deserve their place under the category of emotion, given their inclusion in the corresponding vernacular concept. As a result, emotion encompasses two different kinds of phenomena lacking mechanistic unity. Hence, emotion, as a general category, does not constitute a natural kind.

We can understand Griffiths's analysis of the concept of emotion and his eliminativist thesis as posing a first challenge to emotion research. I propose the following initial construal of this challenge:

Theoretical Challenge (1): Provide a scientifically meaningful theoretical framework that explains all and only the phenomena under the vernacular term "emotion" under the same explanatory resources and under an overarching generic concept of EMOTION.

This construal requires some clarification. First, what characterizes a scientifically meaningful theoretical framework? The literature on the structure of scientific theories is vast, and a proper discussion of this subject matter requires a detailed investigation on its own. Nevertheless, for the purposes of this work, I believe a general working definition of theories can suffice. In my view, Griffiths has in mind an empirical theory, that is, a theory which enables the formulation of empirical hypotheses and that offers clear criteria for its testability. Additionally, I suspect that the theory Griffiths has in mind is also characterized by its systematicity. A theory in this sense is a systematic body of statements which yield predictions and explanations. This working definition allows us to clarify the Theoretical Challenge as follows:

Theoretical Challenge (2): Provide a systematic theoretical framework that provides empirically testable hypotheses and explains all and only the phenomena under the vernacular term "emotion" under the same explanatory resources and under an overarching generic concept of EMOTION.

This construal fares better than the initial construal, but two points needs further discussion. First, how can we determine what are the phenomena under the vernacular term "emotion"?[2] Answering this question is paramount to determining the

[2] I am grateful to Diana Pérez for raising this problem.

scope of our candidate theory. The problem is that the vernacular concept of emotion is a fuzzy concept, and our folk-psychological vocabulary often times cashes out as emotions phenomena that, upon examination, differ significantly from the phenomena traditional theories of emotions have in mind. Consider for example the case of love. For many, love would count as an emotion. Yet, upon close inspection, love is better characterized as a disposition to have certain emotions regarding a loved one (e.g., happiness that they are around, sadness that they leave). Determining whether or not our theory must account of these cases is thus a problematic endeavor. How can we go about defining the scope then?

As in the case of what counts as a scientific theory, we can adopt a working definition of the set. In my view, there are clear cases that the theory must include under its scope: happiness, sadness, fear, anger, and the like. Conversely, there are also other mental (perhaps affective) phenomena that our theory must exclude from the set of emotions: (physical) pain, hunger, difficulty doing complex calculations, executive functions, and so on. By considering these clear cases, we can start approximating the scope of our theory. A theory of emotions must explain phenomena such as happiness, sadness, and fear, and distinguish them from other mental phenomena such as feeling pain, hunger, etc. When it comes to problematic cases such as love, the discussion must be more detailed. However, if a theory fails to explain clear-cut cases or provides a wide construal that includes clear non-emotional cases, we can be sure that theory is not a satisfactory one. Hence, I will stick to clear cases as much as possible in this chapter, and steer clear of more fuzzy, gray cases. In chapters 5 and 6, I will revisit the role of folk emotion concepts in the construction of a scientific theory of emotions. For the time being, I will adopt the aforementioned explananda and evaluate the theories under consideration accordingly.

As for the second point of clarification, as I will argue in Part II, I believe that part of the problem in the emotions and natural kinds debate lies in assuming one particular account of natural kinds, as Griffiths does. There are a number of accounts of natural kindhood and their relation to scientific meaningfulness. Moreover, this heterogeneity of accounts signals a lack of consensus on what natural kinds are and what their connection is to scientific concepts. This introduces an important issue into the debate, as it is unclear what criteria emotions must fulfill in order to qualify as natural kinds and as scientifically meaningful constructs. Even though Griffiths assumes Boyd's (1991; 1999a) homeostatic property cluster (HPC) account of natural kinds, I will discuss this issue in detail later on, when I discuss the issue of natural kinds and their relationship to scientific concepts (see chapter 4). For now, I will leave this issue aside, and grant Griffiths that a satisfactory theory of emotions must formulate kinds in terms of the HPC account.

To analyze the prospects of overcoming the Theoretical Challenge, we must first rehearse Griffiths's argument from the perspective of current emotion research. In other words, Griffihs's original argument, proposed a little more than two decades ago, deserves an update. As it is framed, the argument only considers a subset of current theories of emotions. Therefore, in order to update the challenge, we must

evaluate the best current theories in the market. In order to evaluate them, we must ask whether these theories can successfully explain all emotions and only the emotions under the same framework, and examine the role folk-psychological terms play in these theories.

1.2 Updating the Theoretical Challenge

In order to update Griffiths's argument, two further clarifications are in order. First, I will consider this an update of Griffiths's most general argumentative strategy, leaving aside some of the details of Griffiths's particular line of argument aside. As I understand Griffiths's point, the idea is to examine the best theories of emotions and evaluate whether they are (1) satisfactory theories according to internal criteria of consistency and validity, and (2) they cover clear cut cases of what we would call an emotion and leave out those that we would uncontroversially exclude from this class. In this sense, I follow Griffiths.

Where I don't follow Griffiths, however, is in the choice of the theories I will consider. As I explained above, the theories chosen by Griffiths are no longer the most discussed theories in the literature, hence the diagnosis of his argument as outdated. In my view, the best candidate theories nowadays are Basic Emotion Theories[3], Appraisal Theories, and Psychological Constructionism. The reason why I pick these there theories–or more precisely, families of theories–also responds to Barrett's (2006) review on emotions as natural kinds. In this review, Barrett discusses the two former views, and in later publications (see Barrett, 2012, 2017, 2018a) she defended the latter. Hence, this division reflects the current state of the art and provides a good basis to update Griffiths's argument.

Second, even though much of the current literature has distinguished these three families of theories, this classification is fragile upon close inspection. It is difficult to pinpoint a thesis that separates these three views, and there are a number of arguments that call for integration between theories of different families (e.g. Moors (2017) calls for an integration between dimensional appraisal views and psychological constructionism). I will not discuss these merges in detail at the moment. I will assume this taxonomy of theories is roughly correct and that we can make these distinctions, at least to certain degree. What I am interested in is in the scope of these theories as they have been understood in the literature so far. This also implies that I will leave possible amendments to these theories for now. Later, in subsequent chapters, I will discuss some possible amendments and modifications of these theories. For the purposes of this chapter, however, I intend to present them as they have been presented in the debate and as they form part of the state of the art.

[3] Here I do follow Griffiths slightly, since he does include Basic Emotion Theory in his argument.

1.2.1 Basic Emotion Theory

In general terms, Basic Emotion Theory (BET) aims at explaining emotions in terms of a set of so-called *basic emotions*. The criteria for an emotion to be called 'basic' vary across different versions of BET, and therefore it is difficult to pinpoint an overarching characterization. This also implies that the question of which emotions are basic and which are not (if any, as we will see later) are not consistent among the different theories that fall under the BET umbrella. Given this heterogeneity of theories that fall under BET, I will present BET in a somewhat historical manner. First, I will start with Darwin's work on emotional expression as a precursor of BET. Afterwards, I will present four variations of BET: *Affect Programs Theory* (defended by Tomkins), Ekman's BET (which I will call *Traditional BET*), *Differential Emotions Theory* (defended by Izard), and Panksepp's *affective neuroscience*.

Darwin and *Expression* in BET

In 1872, Darwin published *The Expression of the Emotions in Man and Animals* (1872/2009). Darwin's main contribution in *Expression* is, in a nutshell, the view that emotional expression in humans is not disconnected from expressions in animals, but rather that the former has evolved from the latter. In this sense, Darwin's theory is not a theory of emotions as such, but a theory of *emotional expression*. Nevertheless, this theory did inspire a number of theories a century later, and thus is still an important starting point to understand more contemporary theories of emotion.

As mentioned above, Darwin's main claim is that emotional expression in humans has evolved from expressions in animals. Hence, a proper study of the former involves a study of the latter. This implied a shift in how we thought, not only about emotions, but also about the mind. Historically, dominant theories of the mind viewed animals as devoid from inner lives. At the time of Darwin, Charles Bell, one of Darwin's inspirations, had claimed that emotions were unique to humans and given to us by divine design (Bell, 1844; see also Richards, 2009, p. 114).

Given this historical framework, the shift towards an evolutionary, continuous view of the mind and of emotions was revolutionary. In Darwin's words:

> The community of certain expressions in distinct though allied species, as in the movements of the same facial muscles during laughter by man and by various monkeys, is rendered somewhat more intelligible if we believe in their descent from a common progenitor. He who admits on general grounds that the structure and habits of all animals have been gradually evolved, will look at the whole subject of expression in a new and interesting light. (Darwin, 1872/2009, p. 19)

Studying emotional expression as gradually evolved, as Darwin suggested, implied looking for its origins both in phylogeny and ontogeny. Regarding its phylogenetic origins, Darwin invited us to look at homologous and analogous traits in other species

that may be linked to our own emotional expressions.[4] As for the ontogenetic origins, Darwin proposed studying infants to see the development of emotions, as well as different cultures to see whether there were expressions universal to all human groups. These suggestions, as I will explain later, served as the background of different variants of BET.

Darwin's theory of emotional expression rests in three principles, namely:

Principle of serviceable associated habits Some complex actions are directly or indirectly serviceable under mental states. Whenever the same mental state is induced, there is a tendency formed by habit to make the same movements, even if they are no longer of use.

Principle of antithesis Given the associations between certain movements and a given mental state (as per the first principle), there is sometimes a tendency towards opposite movements when the opposite mental state is induced.

Principle of direct action of the nervous system Some actions are the product of strong nervous system excitation, depending on the connection of nerve-cells and of habit.

Darwin uses these principles to explain the origins of different expressions. For example, he explains the expression of astonishment—characterized by raised eyebrows and open mouth—as formed by habit when our ancestors opened their eyes to expand their visual field and move their eyes quickly to catch the presence of an object in the environment, and hypothesizes that the open mouth comes from disregarding other muscles to focus our attention on the object (which causes the jaw to drop) or as a response to breathe deeper in case we need to flee (see Darwin, 1872/2009, ch. XII). Another example is showing the canine teeth in expressions of anger, which he ties to habits formed in non-human animals to prepare for battle (see Darwin, 1872/2009, ch. X).

Even though many of Darwin's specific claims would be falsified down the line (starting with the inheritance of acquired characters), Darwin's contributions remained an important part of psychology and biology. In the case of emotions and basic emotion theories, Ekman (2009) summarizes Darwin's main theoretical contributions towards BET as follows:

1. Darwin treated emotions as discrete entities, rather than as dimensionally defined as contemporaries like Wundt (1897).

2. Darwin focused on the face as a primary site for evidence about emotions.

3. Darwin thought of emotional expressions as universal, i.e. present in all cultures.

[4] Darwin's theory of expression, even though it was evolutionary, did not make use of the theory of natural selection. Instead, when it comes to expression, Darwin thinks of it as an acquired trait that is inherited to descendants of an organism even if it has no use for its survival (and hence is not a selected trait). See Darwin (1872/2009, pp. 49-54).

4. For Darwin, emotions are not unique to humans but found in many other species.

Regarding the first contribution, Darwin's view of emotions as forming discrete categories is indeed one of the central tenets of different variants of BET, as we will see below. Concerning the second, along with the third contribution, we will see that it is one of the main inspirations of Ekman's later work on emotional expression, which I will present below. Lastly, as for the fourth contribution, this is what I identified as the main theme of *The Expression* and the basis for an evolutionary theory of emotions which BET intends to provide.

As I will show below, all of these themes come up again and again in different variations of BET. In one way or another, every BET is committed to some or all of Darwin's claims as presented by Ekman. It is for this reason that Darwin is commonly known as the precursor of BET.

Affect programs theory (Tomkins)

Ninety years after the publication of *The Expression*, Silvan Tomkins published the first volume of *Affect Imagery Consciousness* (1962/2008, hereafter AIC). In this and the other volumes making up the whole of AIC, Tomkins proposed a theory of affects that updated Darwin's claims and that is still present today in many of our current theories.

Just as Darwin, Tomkins's main interest were not emotions as such. Instead, Tomkins was interested on emotions as leading to motivation. Tomkins's aim in AIC and most of his work is to come up with a satisfactory theory of motivation that is both theoretically sound and empirically productive (Tomkins, 1981/1995). However, Tomkins does come closer than Darwin to proposing a proper theory of emotions. Tomkins's theory of affects does provide, not only the inspiration, but the background and framework for current basic emotion theories.

As Tomkins himself formulates it, the main question he set out to answer throughout his work was: "What do human beings really want?" (Tomkins, 1981/1995, p. 27). He claimed that humans, as any other organism, want to duplicate themselves, that is, to self-maintain and reproduce. To do this, humans have evolved a number of mechanisms that "inform and motivate the individual to incorporate into the organism the raw material from the environment which it must have to remain alive, and informs and motivates the individual to excrete the waste products of the assimilated material" (AIC, p. 18). These mechanisms make up what Tomkins called the 'drive system.'

The drive system, in Tomkins's view, is not sufficient to generate action though. He asks us to consider avoiding being physically hurt. Animals, including humans, have a drive to avoid getting hurt. However, according to Tomkins, this drive itself would not be sufficient to make us avoid getting hurt before we actually do. The reason for this is that the drive system registers being hurt as painful, but the memory of pain itself is not painful. What we need is a way to track how negative previous instances of being hurt have been. This, Tomkins thinks, is done by the *affect system*.

For Tomkins, the affect system amplifies the drives in order to motivate action. Affects in this account are "sets of muscle and glandular responses located in the face and also widely distributed through the body, which generate sensory feedback which is either inherently 'acceptable' or 'unacceptable'." (AIC, p. 135). These sets of facial and bodily responses endow drives with motivational urgency such that we remember and keep track of acceptable or unacceptable stimuli. Without them, drives would not be efficient in keeping us away from harm or forcing us to explore for food and reproduction.

There are three points worth highlighting regarding affect in Tomkins's view. First, Tomkins identifies a set of affects he labels the *primary affects*. These are affects that are activated by different types of neural activity.[5] Which and how many are the primary affects is an empirical question. In Tomkins's words:

> [How many primary affects there are and which are they] is a basic question, primarily biological in nature, that is treated more and more as though it were a psychosocial question. Affect mechanisms are no less biological than drive mechanisms. [...] If each innate affect is controlled by inherited programs that in turn control facial muscle responses, autonomic blood flow, respiratory, and vocal responses, then these correlated sets of responses will define the number and specific types of primary affects. (Tomkins, 1981/1995, p. 58)

The criteria for primary affects are, thus, empirical, biological criteria. It is not a conceptual question which affects turn out to be primary, but an empirical matter concerning our brains and bodies.

Besides proposing biological criteria for primary affects, Tomkins thought that affect is primarily facial behavior. Only secondarily does affect involve bodily, outer skeletal, and visceral behavior (AIC, p. 114). He contrasts his view with the James-Lange (James, 1884) theory of emotion, according to which emotion is the perception of bodily states. For Tomkins, bodily states are part of an affective or emotional state, but only a minor part in comparison to the face. In his view, internal bodily states are slow and gross, whereas the face is rapid and complex (AIC, p. 113). Hence, the primary site for the affects, and consequently the main object of study for a science of emotion, is the face.[6]

Lastly, affects are activated by what he called *affect programs*. An affect program is a set of instructions that "control a variety of muscles and glands to respond with

[5] For Tomkins, there are three distinct classes of affect activators corresponding to different patterns in terms of neural firing density (understood as the product of the intensity times the number of neural firings per unit of time)(AIC, p. 139). These patterns are stimulation (i.e. density) increase, level, and decrease. Each of these patterns contains two or three corresponding affects. Startle, fear, and interest are activated by increasing density patterns; anger and distress, by leveling ones; and laughter and joy, by decreasing ones. For each of these classes, affects are distinguished from one another by the rate of increase or decrease, or by the starting levels of stimulation.

[6] In the next chapter, I will argue against this view. I will claim that evidence on emotional expression is heuristically interesting at best, but cannot decide questions about emotion kinds.

unique patterns of rate and duration of activity characteristic of a given affect" (AIC, p. 135). These affect programs are stored in inherited subcortical structures. They are activated by the presence of a stimulus that has been either innately set up to activate the program (e.g. pain activating a crying response) or that has been associated with the program through learning (e.g. crying because your favorite sports team lost). Each primary affect has a specific affect program that activates it. Consequently, the question of which and how many affect programs are there is also an empirical question.

The idea of there being affect programs corresponding to each type of affect hints at a view that would be central in today's debate, namely, the view that emotions must correspond one-to-one onto some kind of neural structure which instantiates something like an affect program. In the next chapter, I will discuss this idea in detail. For now, it suffices to see where some of the commitments of later versions of BET come from. Most versions of BET expect this sort of correspondence, even using Tomkins's notion of affect programs explicitly. To see this, let us move on to the contemporary versions of the theory.

Ekman's Basic Emotion Theory

Ekman's views derive directly from Tomkins's work. In particular, Ekman's aimed to investigate whether emotional expressions were universal, as Tomkins and Darwin hypothesized. If emotional expressions turned out to be universal, this would presumably provide evidence for the claim that emotions are biologically determined, and thus for an evolutionary theory of emotions along the lines that Tomkins had developed. As we will see below, Ekman's view is premised on the alleged confirmation of such universality of emotional expression. Hence, we can interpret Ekman's BET as an effort to explain and explore the consequences of universality.

Ekman was trained as a psychoanalytic clinical psychologist but a behaviorist researcher. Due to his dissatisfaction with psychoanalysis and his behaviorist leanings, he undertook to investigate behavioral methods of approaching emotional states. He decided to study the face. As he tells his story, his idea of examining the universality of emotional expression was met with resistance from a number of scholars who thought that the issue was already settled against universality and in favor of relativism. Yet, in the midst of such resistance, Ekman found Tomkins, who defended a theory predicting universality and who supported the studies Ekman was to conduct later in his career (Ekman, 1996/2009).

Ekman started by studying emotional expressions across cultures. In the next chapter (2.2.4), I will discuss these experiments in detail. For the time being, I will present the general idea. Ekman and colleagues conducted experiments comparing how emotion expressions were produced and, more importantly, interpreted in different cultures. In his view, these experiments support the hypothesis that there are expressions that are produced and interpreted universally as the same emotions. If this is the case, Ekman argues, there is good reason to believe that there are a number

of emotions present in all cultures. These universal emotions constitute his set of basic emotions.

Since basic emotions in Ekman's construal are independent of culture, they must have biological determinants. In Ekman's view, there are a number of markers that indicate that an emotion qualifies as basic. Ekman (1992) presents the following:

1. Distinctive universal signals.
2. Presence in other primates.
3. Distinctive physiology.
4. Distinctive universals in antecedent events.
5. Coherence among emotional response.
6. Quick onset.
7. Brief duration.
8. Automatic appraisal.
9. Unbidden occurrence.

According to Ekman, the adjective 'basic' in the basic emotions serves to underscore the fact that these emotions form discrete categories, as well as highlighting their biological and evolutionary role. Let us expand on each of these claims.

First, Ekman defends a theory committed to the claim that all emotions form discrete categories. These categories constitute what Ekman calls *emotion families*. An emotion family is a group of emotions that have common characteristics, particularly a common theme. Each of the basic emotions forms one such family. Ekman and Cordaro (2011) present their updated version of the list of (now seven) basic emotion families[7] and their themes as follows:

> **Anger:** the response to interference with out pursuit of a goal we care about [or] someone attempting to harm us (physically or psychologically) or someone we care about.[...]
>
> **Fear:** the response to threat of harm, physical or psychological [which] activates impulses to freeze or flee. [...]
>
> **Surprise:** the response to a suddent unexpected event.[...]
>
> **Sadness:** the response to the loss of an object or person to which you are very attached. [...]
>
> **Disgust:** repulsion by the sight, smell, or taste of something [or] people whose actions are revolting or [...] ideas that are offensive.

[7] Ekman and Cordaro think that there may be more than seven basic emotions. These are only the emotions that they think are confirmed as basic.

Contempt: feeling morally superior to another person.

Happiness: feelings that are enjoyed, that are sought by the person.[...] (Ekman & Cordaro, 2011, p. 365)

According to this list, for example, fear of snakes and dread may fall under the general family of fear reactions. In Ekman's view, all fear reactions share the same expression, physiology, antecedent events, and all of the other markers presented above. Furthermore, they all share the common theme of being a freezing or fleeing response to a threat. The only aspect in which these emotions differ is in the way in which they may be experienced by different subjects. For example, dread may be understood as a particularly intense form of fear, where as fear of snakes may be thought of as a more primitive instance.

Second, Ekman emphasizes the biological and evolutionary role of basic emotions. In his view, emotions have adaptive value and help us deal with what he calls *fundamental life-tasks*. These fundamental life-tasks include situations in which an organism must act rapidly to find solutions to problems they may encounter in the environment. To use the example of fearing snakes, having a rapid reaction to an encounter with a snake may help us avoid death and find refuge promptly.

One interesting but problematic aspect of Ekman's BET is that it doesn't allow for non-basic emotions. For Ekman, there are no non-basic emotions whatsoever. If a candidate to emotion does not satisfy the criteria for basicality, then it does not count as an emotion at all. Rather, it would be cashed out as a mood or as an attitude accompanying one of the basic emotions. This raises the question of the utility of the adjective 'basic', a criticism that has been raised in several occassions (see e.g. Ortony & Turner, 1990). Nevertheless, Ekman has repeatedly defended this use in the sense specified above, namely, as a way to highlight the discrete nature of emotion categories as well as their evolutionary role. This distinguishes Ekman's BET from other forms of BET that allow for non-basic emotions, and will become an important point of discussion later on.

Differential Emotions Theory (Izard)

Besides Ekman, another one of Tomkins's successors was Caroll Izard. Izard's theory resembles Tomkins's and Ekman's in that all of them subscribe to the idea that there are some emotions that are primary or basic. In contrast, however, Izard's view emphasizes the role of feelings in defining emotions. Furthermore, contra Ekman, Izard allows for non-basic emotions, although with some important qualifications which I will present below.

Izard distinguishes between basic emotions and emotion schemas. Basic emotions are "affective processes generated by evolutionarily old brain systems upon the sensing of an ecologically valid stimulus" (Izard, 2007, p. 261). Emotion schemas, in turn, are dynamic interactions between emotion and cognition and may involve complex appraisals (Izard, 2007, p. 265). Emotion schemas are the most common experiences

in adults and older children, and they usually correspond to folk-psychological labels like "fear" or "anger."

Izard's list of basic emotions includes interest, joy, sadness, anger, and fear. These emotions, he thinks, have been shown to have certain properties that warrant their membership to this list. These properties are:

(a) They depend on the perception of an ecologically valid stimulus, but not on complex appraisals or higher-order cognition;

(b) They have more specificity of functions;

(c) They largely derive from bio-evolutionary processes;

(d) They continue to retain relatively more evolutionarily derived features, such as expressive and social signals;

(e) They emerge earlier in ontogeny than emotion schemas; and

(f) They constitute a set of motivational processes important to survival and well-being. (adapted from Izard, 2007, pp. 262-263; Izard, 2011, pp. 371-372)

Basic emotions, as defined by Izard, share a number of features with Ekman's basic emotions and Tomkins's primary affects. All of these are thought of as deriving from evolution and having some adaptive value. Moreover, as we will se below, they are relatively simpler than emotion schemas in the sense that they do not require higher cognitive processes, even though they may interact with them at a later stage in development.

As explained above, emotion schemas are interactions between emotion and cognition. More specifically, an emotion schema is "an emotion interacting dynamically with perceptual and cognitive processes to influence mind and behavior [and are] often elicited by appraisal processes but also by images, memories, and thoughts, and various noncognitive processes and periodic changes in levels of hormones" (Izard, 2009, p. 8). Emotion schemas share the same qualitative aspect with their corresponding basic emotion. For example, an emotion schema based on fear feels the same way as the basic emotion of fear. Such an emotion schema, however, would include other components that the basic emotion does not include. Feeling afraid of failing an exam would involve a complex cognitive process whereby someone judges such a failure as an important loss. Yet, such episode would count as an emotion schema given its complexity.

Just as Ekman's account, Izard's holds that all emotions form discrete categories. According to him, emotions form feeling categories that children learn to capture with the acquisition of language (Izard, 2007, p. 267). Yet, they interact with each other constantly to produce complex behavior. As Izard puts it:

> [...] all emotions are discrete, yet they are highly interactive with each other as well as with cyclical affects (e.g. hunger, sexual arousal) and

> with perceptual and higher order cognitive processes including deliberative thought. (Izard, 2011, p. 374)

Even though Izard emphasizes this sort of general discreteness, it is not entirely clear how Izard thinks discrete emotions are individuated. He thinks of emotions as comprised of neurophysiological, neuromuscular, and phenomenological aspects (Izard, 1971, p. 185). He also stresses the fact that neurophysiological aspects have been shaped by evolution to produce discrete patterns of activation (Izard, 2007). On the surface, it may seem as if emotions were individuated by their neural and physiological patterns.

However, Izard claims that feelings, i.e. the phenomenological aspect, are the key component of emotion. For Izard, each basic emotion has a distinct feeling component which "(a) derives from evolution and neurobiological development, (b) is the key psychological component of emotions and consciousness, and (c) is more often inherently adaptive than maladaptive" (Izard, 2009, p. 3). Second, it is the feeling component that primarily make emotions motivational, leading to different patterns of behavior. Lastly, emotion feelings are specific to each basic emotion, and are shared with emotion schemas based on a given basic emotion. For example, emotion schemas for sadness (e.g. sadness for the loss of a relative and sadness for failing an exam) all share the same feeling corresponding to the basic emotion of sadness. In this sense, it seems that it is feelings, rather than neural or physiological processes, which individuate emotions.

This tension is resolved—at least in part—once we go deeper into Izard's view of conscious feelings. Izard subscribes to a dual-aspect, monistic account of conscious feelings according to which feelings are a phase of neurobiological activity.[8] In other words, conscious feelings for Izard are aspects of neural activity, not different entities that result from the latter. This, of course, invites classical problems regarding monism in philosophy of mind which I will not explore here. Regardless of these problems, these observations make it clearer how Izard might be thinking about emotion individuation: both neurobiological and phenomenological aspects of emotion are candidate criteria for individuation insofar as both are aspects of the same state.

In any case, like his predecessors, Izard claims that emotion feelings, being an aspect of neural activity, are innate and evolutionarily determined. These also constitute the primary motivational component of emotions and are shared across different instances of basic emotions and emotion schemas. Hence, I take it that in Izard's theory, it is feelings that determine whether an episode is an episode of one given emotion or another, and since these feelings are constituted by neural states, we can also individuate emotions at the neural level.

[8] Izard takes this definition from Langer (1967). Langer argues that feelings are phases of neurobiological activity in the sense that they are "modes of appearance" of such activity (see Langer, 1967, p. 21). This position thus amounts generally to a dual-aspect monistic view of conscious feeling.

Panksepp's Affective Neuroscience Perpsective

One last important view among the different variations of BET is Jaak Panksepp's *affective neuroscience*. Even though Panksepp's work did not stem from Tomkins's as Ekman's and Izard's did (Panksepp, 2008), it is still considered among the classical variations of basic emotion views. As we will see below, many of the themes present in these other views are echoed in Pankepp's own account, allowing its membership to this class of theories.

Panksepp's view is premised on the idea that in order to characterize emotions, we must look for homologies in other animals at the neural level (Panksepp, 1998, p. 9). He claims that even though behaviorism banned emotions as scientific constructs, animal research did not, and therefore made progress to find the brain mechanisms responsible for emotional behavior. By importing these findings into human psychology, we can better understand our own emotional lives. This involves integrating the study of behavior, which can be applied to non-human animals, with a neurobiological pespective that tells us about the mechanisms subserving emotional behavioral tendencies.

According to Panksepp, in order the reach a taxonomy of emotions, we must include (1) major categories of human affective experience across individuals and cultures (i.e., folk-psychological categories), (2) a concurrent study of the natural categories of animal emotive behaviors, and (3) a thorough analysis of the brain circuits from which such tendencies arise. Let us examine each of these ideas in turn.

First, Panksepp defends the use of folk-psychological terms in scientific theory. In his view, folk-psychological terms are not gratitious, since they characterize patterns of behavior that can be then studied at a lower level. Put differently, folk-psychological terms serve a heuristic function in emotion research, providing a first glance at a possible taxonomy of affective systems.

Nevertheless, Panksepp is careful to note that some degree of mismatch is to be expected between our folk taxonomies and their underlying biological systems. He recognizes that there is good reason to suspect that folk-psychological terms alone cannot offer a scientific taxonomy. This is partly because it is difficult to define folk-psychological terms in a clear-cut fashion, hence precluding scientists from using these terms fruiftully in generalizations. What is required, argues Panksepp, is a neutral, common ground to define these terms, one that overcomes these difficulties.[9]

In Panksepp's account, it is the neural level that provides such a neutral ground. By defining folk-psychological terms in neural terms, we obtain experimental, neutral criteria on which to ground scientific generalization. Otherwise, he argues, all scientific definitions of emotions would be circular. In his words:

> We cannot say that animals attack because they are angry and then turn around and say that we know animals are angry because they exhibit at-

[9] The account of reconstitution and explication I offer in chapter 5 resembles this position. However, contra Panksepp, I do not endorse the view that we can individuate emotions at the neural level.

> tack. We cannot say that humans flee from danger because they are afraid and then say we know that humans are afraid if they exhibit flight. Such circular word juggling does not allow us to make new and powerful predictions about behavior. However, thanks to the neuroscience revolution, we can begin to specify the potential brain mechanisms that are essential substrates for such basic emotions. When we do that, we begin to exit from the endless rounds of circular explanations. (Panksepp, 1998, p. 13).

Panksepp's agenda then starts with the use of folk-psychological terms and attempts to ground in them in neural systems. How do we go about this? I have already mentioned the general spirit of Panksepp's view: we must study how these terms apply to non-human animal psychology. When we observe non-human animals, we make use of folk-psychological terms to describe some of their behavioral tendencies. We say that cats get angry, that dogs are sad, that rats are afraid, and so on. By noting these uses of language, we can then identify which systems underlie these emotional reactions in other species. The guiding hypothesis is that the mechanisms subserving these functions will correspond to brain circuits that may be homologous to structures in human brains. If this is so, then we can ask which systems are the most evolutionarily ancient and hence more "basic." Panksepp explains this as follows:

> Once we can specify distinct brain systems that generate emotional behaviors, we can also generate biologically defensible taxonomies of emotions. [...] The main criterion here for an emotional system will be whether a coherent emotional response pattern can be activated by localized electrical or chemical stimulation along specific brain circuits, and whether such arousal has affective consequences as measured by consistent approach or avoidance responses. (Panksepp, 1998, p. 14)

With this agenda in mind, Panksepp proposes a scientific definition of emotion. For Panksepp, emotions are "the psychoneural processes that are especially influential in the dynamic flow of intense behavioral interchanges between animals, as well as with certain objects during circumstances that are especially important for survival" (Panksepp, 1998, p. 48). This definition repeats the themes present in other versions of BET: emotions are evolved responses that aid in survival. Furthermore, this definition stresses the importance of brain processes in individuating emotions.

In a more detailed construal of this view, there are a number of criteria to define emotion in neural terms. These criteria are:

1. The underlying circuits are genetically predetermined and designed to respond unconditionally to stimuli arising from major life-challenging circumstances.

2. These circuits organize diverse behaviors by activating or inhibiting motor subroutines and concurrent autonomic-homornal changes that have proved adaptive in the face of such life-challenging circumstances during the evolutionary history of the species.

3. Emotive circuits change the sensitivities of sensory systems that are relevant for the behavioral sequences that have been aroused.

4. Neural activity of emotive systems outlasts the precipitating circumstances.

5. Emotive circuits can come under the conditional control of emotional neutral environmental stimuli.

6. Emotive circuits have reciprocal interactions with the brain mechanisms that elaborate higher decision-making processes and consciousness. (Panksepp, 1998, pp. 48-49).

The list of systems satisfying these criteria constitute what Panksepp calls the "blue ribbon" emotion systems. These are the systems involved in SEEKING, RAGE, FEAR, PANIC. In a later list, Panksepp added LUST, CARE, and PLAY (Panksepp, 2007, p. 286).[10]

One last important idea in Panksepp's account is that emotions form discrete categories that correspond to natural kinds. On one hand, each of these systems constitutes a specific type of emotional response with clear-cut distinctions. Again, this echoes the commitments present in other versions of BET. On the other hand, Panksepp makes explicit his commitment, not only to a discreteness claim, but to the stronger claim that these discrete categories correspond to natural kinds. As he states, given that emotions are individuated by primary processes in the brain, there is a sense in which emotions are bound by something like an essence. More precisely, he thinks of affective natural kinds as akin to species in the sense that projections made across both are made in virtue of their evolutionary history (see Panksepp, 2008). In chapter 4, I will argue that this approach conflates two ways of individuating kinds, namely, via essences and via causal histories. For the moment, it is important to note that Panksepp's view is strongly committed to the idea that emotions do form natural kinds, and thus that emotions can be explained in terms of basic emotion systems.

Problems with BET

At a theoretical level, basic emotion theory is problematic for at least two reasons. First, it is difficult to define criteria for basicality. Different versions of this view use heterogeneous criteria, leading to confusion determining which emotions count as basic. Second, even if this problem were addressed, basic emotion theories still have problems explaining higher order, cognitively complex emotions. Given its focus on evolutionarily ancient, automatic mechanisms underlying emotional responses, basic emotion theories leave out emotions that require more sophisticated mechanisms.

Regarding the first concern, the most influential construal of the problem is offered by Ortony and Turner (1990). In their paper, Ortony and Turner point out that there are at least three notions of basicality at play in traditional construals of the view. As

[10] In the next chapter (see §2.2.1) I will discuss some further implications of this view regarding the problem of whether emotions can be individuated at the neural level.

a result, it is questionable whether the concept of basic emotion is even a useful one. Scarantino and Griffiths (2011) summarize these notions of basicality as follows:

Conceptual basicality An emotion category is conceptually basic ($basic_C$) just in case it occupies the basic level in a conceptual taxonomy.

Biological basicality An emotion is biologically basic ($basic_B$) just in case it has an evolutionary origin and distinctive biological markers.

Psychological basicality An emotion is psychologically basic ($basic_P$) just in case it does not contain another emotion as a component part. (Adapted from Scarantino & Griffiths, 2011, p. 446)

Basic emotions in the conceptually basic sense, as presented above, are those captured by basic-level concepts. This requires an account of concepts that allows for a hierarchy between basic-level and non-basic-level concepts, an account that is not without problems. Even if this account were offered, the problem is that it is unclear whether basic-level concepts refer to basic-level phenomena. For instance, even if conceptually the concept of FEAR were unanalyzable, the phenomenon it refers to might still depend on lower level mechanisms.

Consider an analogy with color concepts such as RED. It might be true that concepts such as RED are unanalyzable in more basic terms. Yet, the phenomenon of color perception, as a psychological and neurobiological phenomenon, depends on mechanisms that are analyzable (e.g. reflection and refraction of light, stimulation of cells in the retina, etc.). For a scientific research program that intends to determine the underlying systems responsible for these responses, the fact that a concept belongs to the basic-level is of limited interest. It might inform theories of psychological concepts, but they might be of little use when studying the underlying phenomena themselves.

In response to these issues, one might adopt a criterion of biological basicality instead. This is the criterion that seems most salient in the theories presented above, with its most explicit formulation stemming from Darwin's views. As Ortony and Turner argue, there are two problems with this criterion. First, as I will explain in detail in the next chapter, evidence for neural or physiological basic mechanisms corresponding to the so-called basic emotions is at least controversial. As far as we know, emotions do not map one-to-one onto mechanisms in the brain or the body, casting doubt on the usefulness of the biological basicality criterion.

Second, Ortony and Turner argue, it is unclear what type of evidence would count to find the biologically basic emotions.[11] Mainly referring to Ekman's account, Ortony and Turner claim that most evidence for biologically basic emotions comes from studies on the universality of facial expressions. Yet, even if we find a set of universal expressions, mapping these expressions onto biologically basic emotions is problematic. Expressions are influenced by cultural factors, to the point that it is difficult to see a "pure" expression of a biologically basic emotion. Ekman tackles this objection by invoking "display rules," socially determined norms whereby subjects supress

[11] In the next chapter, I discuss empirical evidence in depth.

these natural expressions and display their emotions in other ways. According to this argument, whenever we see an expression that does not neatly map onto biologically basic mechanism, we can infer that the pure expression has been suppressed. This move, however, introduces ad-hoc explanations akin to the introduction of epicycles in Ptolemaic astronomy: in the presence of counterexamples (expressions that cannot be easily mapped one-to-one onto biological mechanisms), we introduce a *ceteris paribus* clause into the theory, arguing that these are not normal cases.

Lastly, we find the psychological basicality criterion. Ortony and Turner distinguish two types of psychological basicality, both presented in the formulation above. One is the idea that an emotion is psychologically basic in case we can reduce the eliciting factors of each emotion to a single type. This amounts to explaining, for instance, fear in terms of responses to a threat. This leads, on one hand, to count as basic emotions phenomena that might not be emotions at all. Ortony and Turner mention the case of *courage*. Courage, arguably best described as an attitude than as an emotion, can be described in terms of specific eliciting conditions such as the presence of an undesirable object that is difficult to avoid. Furthermore, it is plausible that all emotions, even canonically non-basic ones such as shame or embarrassment, can be described in the same way, e.g. as responses to the failure of oneself to meet a moral or a social norm, respectively. This risks making the distinction between basic and non-basic emotions meaningless.

This leads us to a final problem with defining clear criteria of basicality. Any interesting criterion of basicality must distinguish between basic and non-basic emotions, otherwise making the category useless. Ekman (1992), however, argues that in his account, all emotions are basic. He argues that he uses the adjective "basic" to stress the evolutionary role emotions play in solving fundamental life tasks. Yet, if this is true, then the adjective itself says nothing. A category is only useful insofar as it distinguishes the cases that fall under it from those that do not, but if all emotions are basic, then basicality is trivial.

In spite of Ortony and Turner's objections, Scarantino and Griffiths (2011) defend these notions of basicality. In their view, there is evidence for conceptually basic emotion, citing Fehr and Russell's (1984) and Shaver et al.'s '(1987) studies. These studies show that some emotion categories such as "happiness" or "pride" exhibit greater prototipicality, shorter names, faster recognition, and other markers of conceptual basicality.

Yet, they recognize that we must distinguish between studying emotion concepts and emotions themselves, in line with Ortony and Turner. Following Scarantino (2012), they distinguish two projects in emotion research. The first, which Scarantino calls the *Folk Emotion Project*, studies emotion concepts as present in our folk-psychological vocabulary. It is in this project where conceptual basicality matters. Nevertheless, the Folk Emotion Project is taken to be a descriptive project with no import on a scientifically tractable account of emotion kinds. This second task is relegated to the second project, the *Scientific Emotion Project*. As Scarantino and Griffiths argue, there may be some mismatch between the categories that the Folk Emotion Project

detects as conceptually basic and the phenomena that the Scientific Emotion Project identifies as biologically and psychologically basic. As a result, they claim that there is an interesting sense of conceptual basicality, but one that we must keep separate from the other types.

Regarding biological and psychological basicality, Scarantino and Griffiths claim that insofar as these are interesting constructs for the Scientific Emotion Project, they can be saved. In their view, objections against these notions of basicality make the mistake of assuming that whatever emotion kinds turn out to be basic in the biological and psychological sense, they must map one-to-one onto folk terms. By distinguishing the aforementioned projects, Scarantino and Griffiths reject this assumption. If we distinguish these two projects, we can keep talking about biologically and psychologically basic emotions even in presence of evidence that they do not map onto folk emotion terms. From this move, they conclude that "from being a mere article of faith, belief in basic emotions still constitutes an empirically promising basis for the conduct of emotion research" (Scarantino & Griffiths, 2011, pp. 452-453).

I am sympathetic to Scarantino and Griffiths's arguments. Yet, as we will see in chapter 5, there are important refinements to be made to Scarantino's distinction between the Folk Emotion Project and the Scientific Emotion Project if we are to find any use for it. To hint at what I will discuss below, I believe that distinguishing these two projects leads to a risk of changing the subject. Assuming that there can be a mismatch between the categories we find in the Folk Emotion Project and the Scientific Emotion Project, we can raise the question: when it comes to the kinds involved in the second project, are we still talking about the same phenomena we call "emotions"? Without criteria to connect folk categories with scientific ones, we run the danger of proposing taxonomies that, while scientifically useful, do not refer to the same set of phenomena, i.e., changing the subject of scientific emotion research. But let us assume for the time being that Scarantino and Griffiths have offered good reasons to believe that a notion of basicality can be offered with this distinction in mind and move on to the second problem regarding basic emotion theory.

Assuming we obtain an account of basicality that is scientifically tractable, it is still difficult for basic emotion theory to explain higher order emotions. The reason for this is that basic emotion theory takes emotions to be automatic, low level processes in the brain and the body. This is especially true for version of BET that characterize emotions in terms of affect programs, such as Ekman's and Panksepp's.

On one construal, affect programs are understood as "pancultural syndromes enabled by inherited biological capabilities" (DeLancey, 2002, p. 3). A syndrome in this sense is a collection of coordinated responses. These syndromes exhibit what Fodor (1983) takes to be characteristic of modular systems. As Griffiths (1997, p. 93) presents them, these features are:

1. The systems' operation is *mandatory* in that it is involuntary.

2. The system is *opaque* to our central cognitive processes, that is, we are not aware of the processes that take place inside the system, only to its outputs.

3. The system is *informationally encapsulated*, i.e., it only has access to information inside the system.

If affect programs are modular systems, then it is hard to see how basic emotion theory could explain cognitively complex emotions. As Griffiths puts it:

> The affect program system creates brief, highly stereotyped emotional reactions. It has only limited involvement with the cognitive processes which control longer-term action. The stimulus appraisal which initiates an affect program reaction is to a large extent informationally encapsulated. The subsequent complex set of actions unfolds automatically, and it is difficult to interfere with these actions voluntarily. There are a large number of emotions which do not conform to this model. In many instances of guilt, envy, or jealousy the subject does not display a stereotypical pattern of physiological effects. In addition, these emotions seem more integrated with cognitive activity leading to planned, long-term actions than the affect program responses. (Griffiths, 1997, p. 100)

Stieg (2007) offers a similar argument. In his view, given the modular structure of affect programs, BET has problems explaining emotions that are deeply connected with social norms. In his view, this is the case of some emotions in non-human animals such as aggression bouts in chimpanzees. These reactions are entrenched in a social context that involves social hierarchies between members of a group. If affect programs are modular, it is unclear how these programs could access information about the social context and other cognitively complex representations that may be required to understand the emotional response. Hence, even in the case of other organisms, BET is not without problems.

There may be some avenues to respond to this objection. On one hand, one may reply, as Ekman arguably would, that these cognitively complex cases are not normal cases of emotions, but rather of emotions coupled with other processes. For instance, one may claim that jealousy is the instantiation of an anger affect program along with thoughts involving a particular situation and social context. However, as I argued above, this introduces *ceteris paribus* clauses into the theory that constitute ad hoc moves. We can generalize this argument to cover other kinds of replies, such as invoking Izard's emotion schemas. In both cases, higher cognitive emotions are not "pure emotions," but some blend or coupling of thep presumed "pure" emotion with other processes.

As a result of these objections, it is clear that basic emotion theory, although an interesting candidate, runs into obstacles when explaining all of the phenomena we call emotions. Either it makes it unclear which emotions count as basic and non-basic, or it excludes from the category instances that we have good reason to include. This is not to say that there are no possible ways to mend the theory though. At the very least, this is to say that as the theory currently stands, it requires important refinements to overcome the theoretical challenge. Let us then consider other views, and see what other problems arise.

1.2.2 Appraisal Theory

Appraisal theories claim that emotions are processes rather than states, contrasting with basic emotion theories. More specifically, they claim that emotions are processes of coordinated changes in different subsystems in an organism in response to the evaluation of a stimulus as relevant to the organism's concerns (Moors, Ellsworth, Scherer, & Frijda, 2013; Scherer, 2005). In other words, these theories see emotions as processes that involve, first, an organism's appraisal of a stimulus as related to its concerns (e.g. survival), and second, the activation of other subsystems that help the organism react according to the appraisal.

Moors and Scherer (2013) hold that a theory of emotions is an appraisal theory in case the theory holds that:

1. Appraisal is a typical cause of emotion or of emotion components; and
2. Appraisal is the core determinant of the content of feelings. (adapted from Moors & Scherer, 2013, p. 135)

On this construal, appraisal theories are those that take appraisal to be at the center of emotion processes. As I will explain below, this can obtain in two ways: (1) appraisal individuates emotions, or (2) appraisal triggers processes involved in emotions. In any of these cases, it is the process of appraisal which determines, at least partially, the ensuing emotion.

So far, appraisal theory has been defined without a proper definition of the appraisal process. On Moors and Scherer's (2013) view, we can define appraisals by their function. So defined, an appraisal is a process which "takes a stimulus (as its input) and produces (as its output) values for one or more appraisal factors (e.g., goal relevance, goal congruence, coping potential, expectancy)" (Moors & Scherer, 2013, p. 137). Depending on which values the function outputs, a set of reactions are triggered in the brain and body to produce the emotion.[12]

Besides their emphasis on appraisal, appraisal theories hold that emotions are built up from a number of components. In spite of the multiple variations among these theories, most agree on the subsystems involved in emotion processes. These subsystems are:

Appraisal component Evaluations of the environment and its relation to the organism.

Motivational component Action tendencies and action readiness.

Somatic component Physiological responses.

Motor component Expressive and instrumental behavior.

[12] On this construal of appraisals, appraisal theories are ontologically neutral regarding the mechanisms underlying appraisals. Moors and Scherer recognize this as an advantage of the view (see 2013, p. 137). While I also take this to be advantageous, I will offer reasons to resist Moors's and Scherer's accounts of emotions below.

Feeling component Subjective feeling.

Among the variants of appraisal theory, Moors (2014; 2017) distinguishes two types. The first are theories that hold discrete emotion categories to be the explananda of emotion research. She calls these *Flavor 1* or *Discrete Appraisal Theories*. The second are those that hold these explananda to be, not discrete categories, but the different subcomponents themselves. These are *Flavor 2* or *Dimensional Appraisal Theories*.

Discrete Appraisal Theories (Flavor 1)

As their label implies, discrete appraisal theories subscribe to the claim that emotions are distinguished by clear boundaries rather than along a set of dimensions. According to these theories, the patterns constituted by the interaction of the different subsystems involved form discrete patterns corresponding to each emotion. In turn, each pattern is individuated and triggered by its corresponding appraisal. Two such views are Lazarus's (1991) and Roseman's (2011; 2013).

In a nutshell, Lazarus's view can be summarized in six principles:

System principle The emotion process involves an organized configuration of many variables: antecedent, mediating process, and outcome of response. No single variable is sufficient to explain the emotional outcome, and all variables are interdependent.

Process principle Emotions are psychological consequences of adaptational struggles (processes), and they demonstrate great variation across time and diverse encounters, because rapid changes in relational meanings are apt to occur at different moments in an encounter and in different encounters, each of which has its own distinctive demands, constraints, and resources (or opportunities).

Structure principle There are stable person-environment relationships that result in recurrent emotional patterns in the same individual.

Developmental principle The biological and social variables that influence the emotions develop and change from birth. The emotion process is not the same at all stages of life.

Specificity principle The emotion process is distinctive for each individual emotion.

Relational meaning principle Each emotion is defined by a unique and specificable relational meaning. This meaning is expressed in a *core relational theme* for each individual emotion, which summarizes the personal harms and benefits residing in each person-environment relationship. The emotional meaning of these person-environ-ment relationships is constructed by the process of *appraisal*. (cited and adapted from Lazarus, 1991, pp. 39, 425)

Put together, these principles amount to the claim that emotions are processes that involve a number of variables, none of which is sufficient for an emotion to obtain, that involve adaptational process that, although variant across different instances

and throughout development, form recurrent patterns that are specific to each emotion, patterns which are described by person-environment relationships (core relational themes) constructed by appraisal.

Unsurprisingly, the role of appraisal is key to this theory. As Lazarus construes it, the appraisal process creates evaluative patterns that individuate each emotion. During appraisal, the organism evaluates specific features of the objects in its environment along a set of variables. There are, according to Lazarus, three *primary appraisals* (goal relevance, goal congruency, and type of ego-involvement) and three *secondary appraisals* (blame or credit, coping potential, and future expectations). Depending on the outcome of these appraisal processes, a set of coordinated responses is triggered. These responses include action tendencies, subjective experiences, and physiological reactions.

One important aspect of Lazarus's view is that even though physiological responses (including those in the brain) are triggered and involved in appraisal processes, they do not map clearly onto folk emotion terms. Lazarus discusses the case of sadness, which can have either a deactivating effect or an activating one (Lazarus, 1991, p. 57). What individuates one emotion from another, in this view, is its relational meaning, the core relational theme it expresses. This contrasts with the views presented above, which expect at least some degree of individuation to occur at the physiological level.

Drawing on the work of Lazarus and others, Roseman also proposed a similar discrete appraisal view. Roseman takes emotions to be *syndromes* in Averill's (1980) sense, i.e. "[sets] of responses that covary in a *systematic fashion*" (Averill, 1980, p. 307, emphasis in original). In this sense, Roseman's view is similar to those presented above under BET. However, this account differs in that it does not include only neurophysiological responses. Instead, the responses Roseman includes are the same as I mentioned above for other appraisal theories (albeit with slightly different names)[13]:

Emotivational (Appraisal) component Evaluations of the environment and its relation to the organism.

Behavioral (Motivational) component Action tendencies and action readiness.

Physiological (Somatic) component Physiological responses.

Expressive (Motor) component Expressive and instrumental behavior.

Phenomenological (Feeling) component Subjective feeling.

According to Roseman, the coordination between these types of responses (i.e. the presence of a syndrome) constitutes a *strategy* to cope with a situation. These strategies are shaped by evolution, aiding in the survival and reproduction of an organism. For instance, fear is the coordination of an appraisal of danger along with physiological, behavioral, expressive, and phenomenological responses that lead the organism to flee or freeze. Moreover, this need not occur consciously or voluntarily, akin to an affect program.

[13] I include the original name in parentheses.

Table 1.1

Roseman's emotion families.

Emotion family	Strategy	Emotions
Contacting emotions	Increase proximity to and/or interaction with impersonal, interpersonal, or intrapersonal stimuli.	Positive emotions.
Distancing emotions	Increase distance from stimuli, thus reducing contact and/or interaction with them.	Distress, sadness, fear, interpersonal dislike, and regret.
Rejection emotions	Move something away from the self.	Disgust, contempt, and shame.
Attack emotions	Move against objects and events in general, against other persons, or against the self.	Frustration, anger, and guilt.

Note: Adapted from Roseman (2011, p. 437).

The different strategies that constitute emotional responses, in Roseman's account, form four contrasting *emotion families*.[14] These families are presented in Table 1.1. Each one of these families includes variation upon the general theme which specify the particular emotion in question.

In Roseman's view, there are categorical differences between each one of these emotions. At a first glance, there are discrete distinctions at the level of strategies. Each strategy constitutes a distinct functional pattern that describes the corresponding emotion. There is also discreteness in terms of appraisals or the emotivational component. An object may be appraised as motive-consistent or motive-inconsistent, but there is no continuum between these two values. Similarly, objects may be appraised as involving either low or high control potential, self-caused or other-caused, and so on (see Roseman, 2011, p. 143). For Roseman, and in contrast to Lazarus, we can even find discreteness in the physiological, expressive, and behavioral levels as well.

Dimensional Appraisal Theories (Flavor 2)

Dimensional appraisal theories, just as discrete appraisal theories, accept a view of emotions in which emotions emerge from a set of coordinated responses that involve the appraisal of a situation. However, in contrast to discrete appraisal views, dimensional appraisal theories claim that emotions are not separated discretely from one another, but rather constitute a continuum along the set of variables involved. For example,

[14] Notice the similarity, again, with BET, particular with Ekman's version. See section 1.2.1 above.

the difference between fear and anger, for dimensional appraisal views, lies in that fear is more of an avoidance emotion but anger is more of an approach emotion. Yet, there are emotions that lie in the spectrum between these two types of action tendencies, precluding a clear cut distinction between the two emotions mentioned.

As Moors (2017) phrases it, the crucial difference between dimensional appraisal theories and their discrete counterpart stems from a difference in what each view takes to be the explananda of emotion research. For discrete appraisal theories, emotion research ought to explain how whole patterns of responses emerge from appraisal processes, as captured by folk emotion terms. Dimensional appraisal theories, in turn, shift the explananda to a lower level. For these theories, emotion research must explain how specific responses emerge from appraisal, regardless of whether they map onto folk emotion terms. In Moors's words:

> [...] instead of trying to explain anger or fear, [dimensional appraisal theories] try to explain the tendencies to dominate, attack, freeze, or avoid, without linking them to anger and fear, and ultimately even, without worrying about whether the components under study are emotional or not. (Moors, 2014, p. 303)

This shift in explananda has important consequences for appraisal theories. Contrary to discrete appraisal theories, which try to map folk emotion terms onto sets of responses and appraisals, dimensional appraisal theories claim that appraisal influences each component independently and that components may influence each other to produce all kinds of reactions. This dynamic, interactive structure between components leads dimensional views to reject the idea that folk emotion categories map onto particular patterns of systemic activation, leading then to skepticism about the scientific status of such categories. In her words:

> Dimensional appraisal theory assumes that there are an infinite number of appraisal patterns that give rise to an infinite number of action tendencies, somatic responses, and experiences, which combine into an infinite number of subsets of emotional episodes. Some of these subsets may fit the profile of vernacular subsets, but most of them do not. (Moors, 2017, p. 7)

Given this skepticism about folk emotion categories, dimensional appraisal theorists claim that we should classify emotions as points in a dimensional space where each dimension corresponds to a component value. For example, an episode of fear may be an constituted by an appraisal that an object is dangerous, a fleeing reaction, increased heart rate, and so on, whereas another may involve a similar appraisal accompanied by a freezing reaction.

An example of this type of theory is Scherer's *Component Process Model (CPM)* (Scherer, 2009a, 2009b). In Scherer's CPM, an event is appraised along a multi-level set of criteria including novelty, pleasantness, discrepancy of expectation, and the like. This multi-level appraisal causes changes in the motivational, motor, and somatic

components. These are then integrated in a central area and, when above a certain activation threshold, reach consciousness in terms of a feeling with its own *qualia*.

According to Scherer's CPM, since appraisal is multi-leveled and emotions divide into several components, there may be infinite different combinations of appraisal values and activation patterns. As a result, "emotion differentiation is the result of the net effect of all subsystem changes brought about by the outcome profile of the [stimuli evaluation checks]" (Scherer, 2009a, p. 1314). Additionally, these subsystem changes influence one another in a number of ways, producing recursive structures and feedback effects that affect the net outcome that determines the emotion.

Two aspects of this model are worth highlighting. One is that, according to it, there are as many emotions as there are possible net effects of subsystem changes. These net changes, in turn, depend not only on the appraisal itself but also on the activation patterns of other components and their interaction. As such, emotions are best described in a high-dimensional space where each dimension corresponds to each component's value. Naturally, appraisal plays a causal role in triggering different components, but it does not completely determine the resulting emotion.

A second important aspect of Scherer's CPM is that, according to the author, it does not constitute a 'natural kind' model of emotion. given its rejection of Discreteness and its expectation that emotions form fuzzy sets determined by the net action of different subsystems, the CPM does not commit itself to the idea that emotions form specific, consistent, or unique kinds of phenomena. Rather, emotion is taken to be an emergent phenomenon. In this regard, Scherer's CPM distances itself from basic emotion theories and discrete appraisal theories.

Problems with appraisal theories

Appraisal theories offer a promising, interesting view. In contrast to other theories, appraisal theories highlight the different aspects involved in emotions while providing a framework that makes visible what is common to emotional responses. Furthermore, given the compatibility of appraisal theories with functionalist views, problems regarding heterogeneity in the underlying mechanisms of emotions are not as pressing as for other views like BET.

Although I am fairly optimistic about appraisal theories, there are some obstacles it must hurdle if it is to overcome the theoretical challenge. First, it is unclear whether we can map appraisals onto emotional reactions one-to-one. Frijda and Zeelenberg (2001), for instance, claim that not all fear is caused by the anticipation of danger, but can be elicited by novelty or unexpectedness. More recently, Kuppens, Van Mechelen, Smits, De Boeck, and Ceulemans (2007) have shown that anger can be triggered by a variety of appraisals, including appraisals of accountability and unfairness which may or may not be present in a given instance. This affects especially discrete accounts which hold that emotions are individuated in terms of appraisals.

One possible response is to claim that the link between appraisals and emotions is a conceptual rather than an empirical one. According to this line of argument, fear is just that reaction that occurs in the anticipation of danger; other reactions such as

those elicited by novelty or unexpectedness might be related, but do not technically do not constitute instances of fear. What is contingent on this view is the relation between appraisals and reactions such as physiological or phenomenological responses. It is here where emotions are home as objects of scientific study.

This way of replying to the objection above shows a second problem appraisal theories face. If the relation between emotions and appraisals is conceptual, then many of the presumed empirical results confirming the theory are trivial. Smedslund (1992), commenting on Frijda's (1988; 2007) account of emotions, puts forward this sort of argument. In his view, Frijda's account, which cashes out emotions in terms of the awareness of relationships of importance to an individual's concerns, makes pseudoempirical claims. The main reason is that the link between the relationships of importance and emotions is conceptual, hence noncontingent and not suitable for empirical verification. Taking the example above, fear is defined as the awareness of something as dangerous to us, for instance. This makes any finding of this relation trivial.

Smedslund's argument can be generalized to appraisal theory, or at least discrete appraisal theories. By relying on appraisal for individuation, the theory risks collapsing into triviality. I do not think this is a finishing blow against the view, however. In later chapters (see chapter 6), I will argue that we can formulate a theory of emotions where relations between the organism and its environment can be empirically interesting while playing a central role in the individuation of emotions. For now, it suffices to point out that in order to meet the theoretical challenge and offer a scientifically sound framework for emotions, appraisal theories must clarify the relations between appraisal and emotions so as to avoid triviality.

Besides these worries, which are more pressing for discrete appraisal theories, there are other concerns that affect dimensional appraisal views. One is that for dimensional appraisal theories, there are potentially infinite types of emotions, as many as there are possible combinations of appraisal and subcomponent values. This makes any classification of emotions arbitrary, since what distinguishes one emotion from another is nothing regarding the projectibility of an emotion category. In other words, what makes each emotion—understood folk-psychologically—that emotion is nothing beyond an arbitrary border drawn between points in a continuous space of infinite emotions. Such arbitrariness is not only problematic in itself, but leads to deeper issues.

One issue with the arbitrariness of emotion taxonomies for dimensional appraisal views is that, methodologically, emotion categories have proven fruitful in the scientific study of emotions. Elsewhere I have argued for this claim, along with Eickers and Prinz (Eickers, Loaiza, & Prinz, 2017). Discrete categories have been used in a number of successful studies and, as I will argue in detail in the last chapter, they can be accounted for functionally. If this is so, the skepticism that characterizes dimensional appraisal theories is unwarranted.

But even leaving this issue aside, dimensional appraisal theory's skepticism has a deeper consequence for our current discussion. If emotion categories are rendered

arbitrary by said skepticism, then there is a sense in which dimensional appraisal theories are not theories of emotion at all. This is because, given the arbitrariness of emotion categories, there is no interesting sense in which relations between appraisals and reactions explain what we call "emotions." As with Scarantino's Scientific Emotion Project above, dimensional appraisal theory risks changing the subject by drifting too far away from what emotion resarch ought to explain. Hence, the theory is unable to overcome the Theoretical Challenge, since it does not offer a framework to study emotions proper. Put differently, if differences between emotion categories are left unexplained, the theory does not meet the Theoretical Challenge. It does not explain the phenomena under the vernacular term "emotion," as it renders taxonomies of emotions arbitrary.

1.2.3 Psychological Constructionism

Psychological constructionism about emotions is the claim that emotions are not atomic entities, but rather are composed of other building blocks, as it were. More specifically, constructionists claim that emotions are not ready-made phenomena independent of ourselves, waiting to be felt by subjects and perceived by others. Instead, they submit that emotions are actively built by us every time we feel them and act on them.

Constructionists reject many of the claims basic emotion theories subscribe to. These theorists think that emotions have no neural or bodily fingerprints to which they correspond one-to-one. Moreover, they claim that there is no interesting sense in which emotion categories could map onto bodily or neural states, given that one such state can instantiate a variety of emotions depending on external variables such as context and history. Instead, constructionists think that this variability is not only established but normal. As a result, constructionists reject any form of discretness, universality, and innateness of emotion mechanisms. Russell (2009) summarizes this disagreement in the following claims:

- There are cultural differences in all known aspects of emotion.
- Different languages lack a one-to-one correspondence between emotion terms.
- Theories based on traditional assumptions have not led to increased precision of terms. Each term lacks inclusion and exclusion rules.
- Basic emotions rarely occur alone, and yet no accepted theory of how they co-occur or blend has been developed.
- Failure to find convincing evidence that emotions produce "facial expressions of emotion".
- Failure to find convincing evidence of a unique pattern for each emotion in the autonomic nervous system.

- Failure to find separate factors corresponding to basic emotions in studies of self-reported emotional experience.
- Failure to find a class of behaviour common to instances of a given emotion.
- Dissociation rather than predicted associations among manifest components. (Russell, 2009, p. 1261)

Besides rejecting basic emotion theories, constructionists also reject some claims coming from appraisal theories. In contrast to the latter, constructionists think that emotions require the active participation of an agent in order to obtain. For appraisal theorists, an emotion obtains if there is coordinated activity from the corresponding subsystems, but this is not dependent on the agent's psychological history or social context. Constructionism, on the other hand, does claim that our own learning experience, categories and social contexts determine which emotion is instantiated.[15]

Psychological constructionism may be divided into two broad categories. One on hand, there is the view that the construction of an emotion requires an act of categorization which requires the possession of emotion concepts. This is *Conceptual Psychological Constructionism.* On the other hand, there is the opposite view that such act of categorization can occur without concepts as long as the necessary ingredients obtain. This is *Non-conceptual psychological constructionism.*[16] I will present non-conceptual psychological constructionism first, since conceptual psychological constructionism builds on the former.

Russell's non-conceptual psychological constructionism

The main proponent of non-conceptual psychological constructionism is Russell (2003, 2009, 2015). Along with other psychological constructionists, Russell defends a skeptical view about the utility of folk emotion categories such as 'fear,' 'anger,' and the like. In his view, folk emotion terms have no scientific value, as they have lead to the assumptions that each term should map onto a specific mechanism in the brain and the body, and that these mechanisms should be distinct from each other as these terms are.

Given this rejection of folk emotion category, how can we understand emotions? Russell proposes to think of emotion as composed by serveral components and experiences that get categorized in a particular way. These components include somatic changes, expressive behavior, and other components that other theories highlight. However, and more importantly, there are four components that are central to Russell's account.

First, Russell claims that at every moment of our waking time, we are constantly in a state of pleasure or displeasure, i.e. valence, and of activation or deactivation, i.e. arousal. The combination of these two states into a single two-dimensional construct

[15] Additionally, given that constructionists deny discreteness, they also differ this respect from discrete appraisal theories.

[16] I borrow the labels from Scarantino (2015).

is what Russell calls *core affect*. Technically speaking, core affect is "that neurophysiological state consciously accessible as the simplest raw (nonreflective) feelings evident in moods and emotions" (Russell, 2003, p. 148) . It is important to note that core affect is general to all affective experience, including moods but also attitudes and other phenomena we might include in this category. For this reason, and as we will see below, there is no mechanism or system that distinguishes emotions from other affects.

Core affect, Russell states, is universal and simple (Russell, 2003, p. 148). It is a continuous monitoring of one's own state and can exist without us categorizing it or interpreting it as part of some affect. It is not dependent on cognitive processing, hence not being dependent of any concepts we may possess. Lastly, it can be manipulated by chemical means such as stimulants or antidepressants.

Second, Russell points out that some objects appear to us as possessing certain qualities related to our affects. We find objects pleasant or unpleasant, upsetting or calming, and so on. What makes these objects pleasant or unpleasant, upsetting or calming, etc., is their capacity to change our core affective state. A pleasant object puts us in a pleasant state; a calming object in a low arousal state. These qualities, Russell holds, are part of our representation of these objects, and as such they are perceived as external to us. These he labels *affective qualities*.

Third, when we perceive an object as having some affective quality and as a consequence perceive changes in our core affective state, we may link these two events and interpret them as causally related. Such a combination of affective qualities and core affect is called *attributed affect*. When we experience an attributed affect, we experience an object as changing our core affect and as worthy of our attention and subsequent behavior. This explains emotional utterances such as "I am afraid of a snake" as cases where we perceive a snake as changing our core affect into a high arousal, low valence state.[17]

Lastly, when we perceive an attributed affect, we may categorize that affect as an instance of an emotion category. Hence, we may experience the affective process as one of 'fear,' 'anger,' 'happiness,' etc. At this point we have what Russell calls an *emotional meta-experience*. These emotional meta-experiences require concepts, but they are only a component of an emotional episode. Furthermore, our emotional meta-experiences may categorize an affective episode in the wrong way or not obtain at all. To use Russell's own examples, we may feel jealous and deny feeling jealous. Also, according to Russell, children may have affective experiences without emotional meta-experience, given that they do not have the concepts to categorize their own states (Russell, 2003, p. 164).

Given these components, Russell summarizes a "prototypical emotional episode" as follows. First, there is some event; for instance, we see an insect. This event or its object, in this case the insect, is perceived to have some affective quality, say we find it disgusting or scary. This automatically shifts our core affective state from low arousal

[17] Notice that at this point, Russell is close to formulating a theory of the intentionality of emotions. He does not offer such theory though.

to high arousal, and neutral valence to negative valence. As these changes happen, we attribute the changes to the object, the insect, an enter in an attributed affect. At this point, we may immediately jump away or scream. This involves changes in our physiology and expression, which in turn create other subjective experiences as we feel these reactions happening in our body. Lastly, if we have a concept of disgust or fear, we may categorize this experience and engage in an emotional meta-experience.

Russell's constructionism attempts to integrate ideas from dimensional theories (e.g., dimensional appraisal theories) and categorical analyses (e.g., basic emotion theories). In his view, core affect by itself cannot explain differences between emotion categories, a problem I explained above when discussing dimensional appraisal theories. As he states it:

> [...] by themselves, pleasure and arousal do not fully account for most emotional episodes. Specifically, I acknowledge that my own dimensional model of emotion (Russell, 1980) does not provide a sufficiently rich account of prototypical emotional episodes. For example, that model fails to explain adequately how fear, jealousy, anger, and shame are different and how observers can distinguish them. The dimensional perspective must be integrated with the categorical perspective [...]. (Russell, 2003, p. 150).

How, then, are emotion categories distinguished from each other? In other words, what individuates emotions? According to Russell, emotions are individuated in virtue of mental scripts that define an emotion category. Let us unpack this claim. First, for Russell, emotion categories have a prototype structure (Russell, 1991). This means that what counts as a member of the category does so because it resembles a prototypical instance. For instance, an episode of sadness counts as such if it is similar enough to a prototypical episode of sadness. Thus, the question concerning emotion individuation becomes a question about identifying prototypes.

According to Russell, these prototypes constitute mental scripts, sequences of subevents that typical instances of an emotion follow. For example, Russell (1991) presents the following analysis of the script of anger:

1. The person is offended. The offense is intentional and harmful.
2. The person is innocent. An injustice has been done.
3. The person glares and scowls at the offender.
4. The person feels internal tension and agitation, as if heat and pressure were rapidly mounting inside. He feels his heart pounding and his muscles tightening.
5. The person desires retribution.
6. The person loses control and strikes out, harming the offender. (adapted from Russell, 1991, p. 39)

On this analysis, instances of anger qualify as such because they resemble this prototypical pattern, even if exceptions or deviations are at play. For instance, cases of anger where the offended person does not strike out, or where there is no real injustice but only perceived injustice, also count as instances of anger in virtue of their resemblance to this prototypical script.

For Russell, this analysis integrates dimensional and categorical claims about emotions. It concedes to dimensional analyses the idea that emotions are constructed of domain-general responses such as core affect. However, it also makes room for an explanation of how emotional categories come about, namely, but the grouping of patterns of affective responses in terms of resemblance to prototypes. Below I will argue that even though this approach is promising, non-conceptual psychological constructionism still has problems explaining emotion taxonomies in a non-arbitrary way. Before I discuss this, let us present the other variant of psychological constructionism, namely, conceptual psychological constructionism.

Conceptual psychological constructionism (Barrett)

Barrett (2017; 2018a) is the main proponent of conceptual psychological constructionism. Similar to Russell, Barrett claims that emotions are constructed from more basic psychological building blocks that are not themselves emotional. In contrast to Russell, however, Barrett holds that concepts are necessary to construct emotions.

In early writings, Barrett expressed skepticism regarding the scientific status of folk psychological categories. In an influential review (Barrett, 2006) and subsequent writings (Barrett, 2012, 2018a; Barrett et al., 2007), Barrett argued that most of the tradition in emotion research has presupposed what she calls the "natural kind view" of emotion. According to the natural kind view, emotions are discrete categories that correspond to folk categories such as 'fear,' 'anger,' and so on, and that map onto specific, unique, and consistent regions or circuits in the brain, as well as distinct physiological states. In her review, she presents evidence against this view, showing that empirical research has found variation for each of these categories. Importantly, she invites us to take variation seriously, claiming that this observed variation is not a mere methodological artifact awaiting for correct methods of investigation, but rather the norm. In other words, she considers variation to be established and evidence for a different account of emotions.[18]

Her positive view is based on the idea that emotions are constructed by acts of categorization by means of concepts. Hence, she has referred to her view as the "Conceptual Act Theory" (Barrett, 2014) in the past. Later she coined the name "Theory of constructed emotion."[19] According to Barrett, an emotion is constructed in the sense that our brains create categories based on our experiences that aid in predicting behavior. Before a child acquires emotion concepts, the child is exposed

[18] The claim that variation is established is what I call the Variability Thesis, which I discuss in detail in the next chapter (see chapter 2).

[19] Elsewhere, Barrett has explained that the change in name was due to editorial reasons rather than substantial changes to the view. See Barrett (2018b)

to different responses from its brain and body, as well as to the behavior of others around it, which it learns to categorize given certain similarities. These similarities then form a concept which the child learns to map to a label in a given language such as 'sadness,' 'Traurigkeit,' 'tristeza,' and so on.

For Barrett, concepts are a necessary part of the construction process. Without concepts, all we would have is access to a pattern of sensory signals and behavior. She proposes an analogy with experiential blindness. In experiential blindness, an agent sees a stimulus which they cannot categorize. Once the agent acquires a concept of what the stimulus is, the agent now sees the stimulus as a token of that concept. A similar situation applies to emotions in Barrett's view. Emotions may involve bodily signals and contextual cues. But, contrary to other theories, Barrett thinks that these only constitute elements of emotion once we have acquired the concepts required to categorize them. Before acquisition, we are experientially blind to our own emotions, just as we were experientially blind to the bee. This process of construction is what creates emotions every time we have an emotional episode. This aspect of Barrett's theory contrasts with Russell's view, in which components are part of the ensuing emotion regardless of categorization.

In the case of the bee, we have black spots as the basis on which our brain constructs an image that we later perceive as an image of a bee. What is the basis on which our emotions are constructed? An important ingredient in this view is *core affect*, just as in Russell's account. For Barrett too, we are constantly in some state of core affect which can be described in terms of valence and arousal. Yet, Barrett adds more to the recipe. On one hand, Barrett thinks of affective feelings as predictions. In this picture, affective feelings are the ways in which our brains detect our core affective states and formulate predictions based on that information. These predictions may involve hypotheses about dangerous things in the environment that later construct an instance of fear, or about winning an unexpected prize which leads to the construction of happiness.

Other ingredients may also fall under the predictions our brain does to construct emotions. For example, seeing someone cry may inform our brain that the other is in distress, and hence construct an instance of sadness. Other information can include reading a story about someone in a difficult situation with no way out, or watching a movie we consider sad. As we receive that information, our brains predict instances of sadness and, if we have the adequate concept, constructs an emotion episode.

Barrett's view rejects discreteness from the outset. In this account, emotions are not discretely organized but rather continuous all throughout. In Barrett's terms, "instances of emotions are *momentary snapshots of continuous brain activity*, and we *merely perceive* these snapshots as discrete events" (2018a, pp. 36-37; emphasis added). However, in contrast to Russell, Barrett cannot be said to be committed to a dimensionality claim either. The reason is that for Barrett, even though core affective states may be described dimensionally, emotions are much broader than core affect. Hence, strictly speaking, emotions do not lie in a continuous space of core affective states.

What is clear though is that Barrett's view strongly rejects claims for consistency and specificity of the underlying mechanisms of emotions. A mechanism (response or set of responses) is said to be consistent for an emotion category if it is present in all instances of that emotion. For example, hypotheses mapping fear to activity in the amygdala would hold that amygdala activity is consistent for fear. A mechanism is also said to be specific to that emotion if it is only linked to that emotion and not others. For Barrett, evidence shows that emotions are heterogeneous at the neural, physiological, behavioral, expressive, and phenomenological levels. Because of this, claims for consistency and specificity are taken to be falsified.

Barrett's account, however, goes beyond merely claiming that consistency and specificity are false claims. In Barrett's view, it is not only that they are false, but that theoretically we should not expect either of these claims to be true. Instead, a theory of emotions should be compatible with variability and explain how emotions could arise even in the absence of consistent and specific mechanisms. This is the reason why she summarizes one of the main tenets of her view under the motto "variation is the norm" (Barrett, 2018a, p. 23).

Problems with psychological constructionism

When it comes to meeting the theoretical challenge, psychological constructionism fares no better than other views considered above. For defenders of constructionism, this might appear not to be a problem at all, since they hold that the Theoretical Challenge, so construed, is not a challenge worth considering. On this interpreation, psychological constructionism is not committed to a stong distinction between emotions and other affective phenomena. Instead, they argue that emotions and other affects are produced by the same domain-general mechanisms underlying valence and arousal. As a result, the challenge of proposing a theory that would explain emotions and only emotions is not a challenge they would undertake to overcome.

This interpretation applies particularly to earlier versions of psychological constructionism. For example, on Russell's account, there may be affective responses without the construction of an emotion in cases where what he calls an emotional meta-experience does not obtain. Moreover, when an emotional meta-experience obtains, it is only because the subject interprets their core affective state in a particular manner, but not because of any specific mechanism underlying their responses. Hence, distinguishing emotions and other affects does not constitute an important matter worth elucidating. As he explains in his own words:

> As a scientific term, *emotion* is separated from *not-emotion* (or *fear* from *not-fear*, etc.) in a precise way only by stipulation. I therefore suggest that we begin to move away from these everyday words as technical terms. If we do so, we would not need to define *emotion* prescriptively; that is, we would not need to stipulate a scientific definition that states the boundaries of the set of events to be explained, because *emotion* is given no scientific work to do. For example, my concept of core affect is not defined in

> terms of emotion. On the other hand, *emotional episode* does contain the word *emotional*. Emotional episodes must be explained, because all human episodes must be explained, and not because emotional episodes are qualitatively different from nonemotional episodes, whatever those might be. In other words, the goal of psychological construction is an account that includes, but is not exclusive to, all events labeled as *emotional*. So, in my account of emotional episodes, emotion functions something like a chapter heading and does not function as a technical term. (Russell, 2015, p. 205)

Nevertheless, more recent versions of constructionism, especially Barrett's TCE, do seem to attempt to offer a theory of emotions that would explain all emotions under the same framework. It is unclear though whether this constitutes an attempt to overcome the Theoretical Challenge. Barrett's view still holds that emotions are constructed when we interpret activity in domain-general systems as belonging to an emotion category. This may be interpreted as a way of explaining what is particular about emotions, namely, that they are interpretations using a specific set of concepts we may call "emotional concepts."

If we interpret Barrett's TCE, and perhaps other forms of constructionism, as attempting to overcome the Theoretical Challenge, we find at least three problems. First, the notion of *core affect* that lies at the heart of constructionism is questionable from a naturalistic perspective. As I explained above, core affect is defined as a neurobiological state of arousal and valence. We could concede that arousal might come to be naturalized in terms of physiological responses such as heart rate, electrodermal responses, and the like. However, naturalizing valence is not a straightforward task. It is unclear what would count as a valenced states merely in neurobiological terms. At best, we could propose an analysis of valence in terms of harm/benefit relations between an organism and objects in its environment as concerning its survival, but this hardly suffices to explain differences between emotional states. For instance, what makes pride a positive emotion is not merely that objects of pride are beneficial to the survival of the organism. Similarly, it is difficult to see how shame could come to be a negative emotion in terms of harmful objects in the environment. Without a clear account of core affect, the theory cannot propose a naturalistic, scientific theory.[20]

Second, as with dimensional appraisal theory, constructionism makes emotion categories relatively arbitrary. On these views, there is no reason why we should distinguish fear from anger, anger from happiness, and so on, besides conceptual truths that do not respond to any epistemic aims. In other words, according to constructionism, emotion categories do not serve any scientifically interesting classificatory purpose such as separating different kinds of responses. Instead, emotion categories are constructs determined by our social frameworks. This arbitrariness precludes these categories from being used in generalizations and inductive inferences, thus banning them from scientific theories.

[20] Elsewhere I have argued for this claim. See Eickers et al. (2017).

Lastly, psychological constructionism submits that emotions are wildly heterogeneous phenomena, opposing what Barrett has called the natural kind view of emotions. Yet, in my view, constructionists appeal to variability without an adequate understanding of this claim. This precludes them from establishing it empirically, since there are no clear confirmation criteria for this thesis. Additionally, in spite of there being some evidence in their favor, constructionists overstate empirical findings to fit their view and assume them as true from the outset. As I will discuss in detail in the next chapter, it is not the case that evidence unequivocally supports constructionism. As a result, without a clear account of what variability means, constructionists cannot get their theory off the ground, nor can they assume at the theoretical level that emotions are as heterogeneous as they take them to be. In other words, variability must be a claim to be established empirically, but constructionists lack an account of variability that would allow them to do so.

1.3 Prospects to overcome the Theoretical Challenge

Let us go back to the Theoretical Challenge as presented at the beginning of this chapter. The Theoretical Challenge calls for a systematic theoretical framework that can explain all and only the phenomena under the vernacular term "emotion" in an empirically interesting and conceptually sound way. As I have shown in this chapter, none of our best theories in psychology and neuroscience overcomes this challenge. Each of these theories or families of theories have problems that, at least as they stand in the current literature, require addressing before they can meet the challenge.

To generalize from the arguments I have presented against each theory, we can see two main problems current scientific theories of emotions face. First, many of the theories presented above rely on constructs that are dubious at best. These include the notion of basicality in the case of basic emotion theories, the construct of appraisal as a means of emotion individuation in appraisal theories, and the construct of core affect in psychological constructionism. These categories are not well-defined, which makes the theories they support conceptually problematic. Without addressing these issues, these theories cannot provide a sound theoretical basis on which to draw empirical hypotheses. As a result, none of these theories meets the challenge above.

A second problem which is present in some of the theories above is that they render distinctions between emotions arbitrary. As we will see in chapter 5, this is a major issue with dimensional theories in general, which include dimensional appraisal theories and psychological constructionist theories. The main reason why this is problematic is that it risks changing the subject, as the phenomena under the vernacular concept "emotion" do involve distinctions that a scientific theory of emotions should be prepared to explain. By rendering these distinctions arbitrary, these theories cannot explain what distinguishes emotions from other affective or even mental phenomena, neither can they address the question of how emotions differ from each other besides our conceptual framework. Given these problems, dimensional theories cannot overcome the Theoretical Challenge.

Seeing the discussion in this chapter as an update of Griffiths's arguments, one could think this suggests that emotions should be eliminated from scientific discourse, as Griffiths does. Recall that Griffiths argued that if none of our best theories can account for emotions under the same theoretical framework, there is a sense in which emotions do not exist as they do not form a natural kind. Consequently, claims Griffiths, we ought to eliminate emotion concepts from scientific theories.

While I do not share Griffiths's conclusion, as will be clear at the end of Part I (see chapter 3), I believe these problems should be taken seriously. The fact that none of the current theories in psychology and neuroscience meet the Theoretical Challenge means that we require a new theory of emotions at the very least. If such a theory is impossible in principle, then there is good reason to opt for an eliminativist view.

Before I discuss the prospects for eliminativism in detail, let us go over the other side of the problem, what I call the Empirical Challenge. In the next chapter, I will examine the prospects of finding empirical evidence that would support a theory of emotions that construes them as homogeneous, projectible categories. I will argue that evidence suggests an important degree of variability for each emotion category and across the general category of emotion altogether. Yet, for this claim to support eliminativist conclusions, a more precise account is necessary.

After I analyze the Variability Thesis, as I will call it, I will revisit eliminativist arguments and argue that they are unwarranted. This will serve as the basis for the Part II, where I defend a revisionist position according to which a new theory of emotions is required. I will offer some criteria which, in my view, a scientifically interesting theory of emotion must satisfy, and I will attempt to answer the question of how a satisfactory theory of emotions might look like.

Chapter 2

The Empirical Challenge

The Problem of Variability

In this chapter, I will discuss the *Empirical Challenge* in terms of what has been called in the literature the Problem of Variability (Scarantino, 2015). The Problem of Variability stems for the thesis that emotions are naturally disjoined phenomena (under some criteria of unity which I will discuss below). Call this the Variability Thesis (VT). If this thesis is correct, it would presumably follow that emotions do not form natural kinds, hence precluding the formulation of a proper unified scientific theory.[21]

Past failures to reject VT constitute a source of skepticism among a number of emotion researchers. It is the backbone of Barrett's (2006) attack on the claim that emotions are natural kinds. Hence, a tractable scientific theory of emotions, in the sense demanded by the Theoretical Challenge, would in principle require rejecting VT. In other words, if a unified theory of emotions in terms of a common set of explanatory resources is available, it must be able to show how emotions form a unified category at an empirical level. This is what I call the *Empirical Challenge*.

In order to offer a detailed account of the Empirical Challenge, we must first analyze the Variability Thesis. By doing so, we can be clear on what is exactly what is at stake in the claim that a theory of emotions must enable us to reject VT. This would provide an idea of what precisely is the demand involved in the Empirical Challenge and its consequences for scientific research on emotions.

This chapter thus offers an analysis of VT which allows a clarification of the Empirical Challenge. In the first section, I will present VT and how it has been presented in the debate. Then I will examine some conceptual problems involved in these previous formulations of the thesis, and propose an analysis that ameliorates some of these issues. I then identify as one of its core problems the issue of the individuation of patterns of responses, which requires a discussion in terms of each domain relevant to the investigation of emotions. Hence, I proceed to discuss each empirical approach to individuate patterns in each of the relevant domains, and close with a proposal

[21] Parts of this chapter were published in Loaiza (2020).

as to which patterns empirical research should focus on and suggest some ways of individuating said patterns.

2.1 The Variability Thesis

The Variability Thesis (VT) can be presented under the following working definition:

Variability Thesis (VT) Emotions are naturally disjoined phenomena.

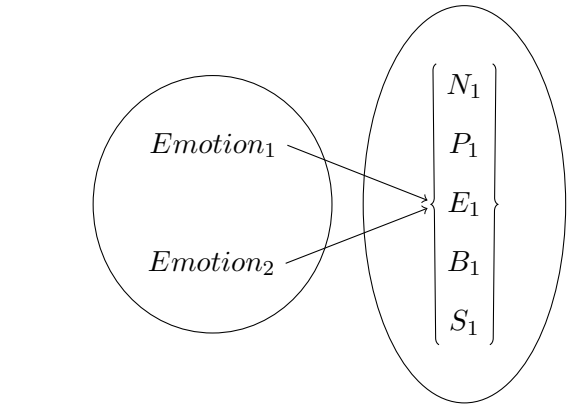
In the current literature, VT is taken to stem from previous efforts to find homogeneous processes corresponding to emotion categories. As Scarantino (2015) and others (Barrett, 2006; Prinz, 2004) formulate it, it is a thesis that affects primarily basic emotion theories, given that these theories (in at least some of its incarnations) expect hardwired circuits and discrete physiological patterns that map onto emotion categories, hence expecting homogeneous patterns for each individual emotion. However, this claim is not exclusive of basic emotion theories. Barrett (2006) argues that it also affects discrete appraisal theories, since these theories also expect certain degree of specificity and correspondence. In any case, VT is a problematic thesis for any theory that expects specificity and correspondence.

Scarantino presents VT as constituted by two theses:

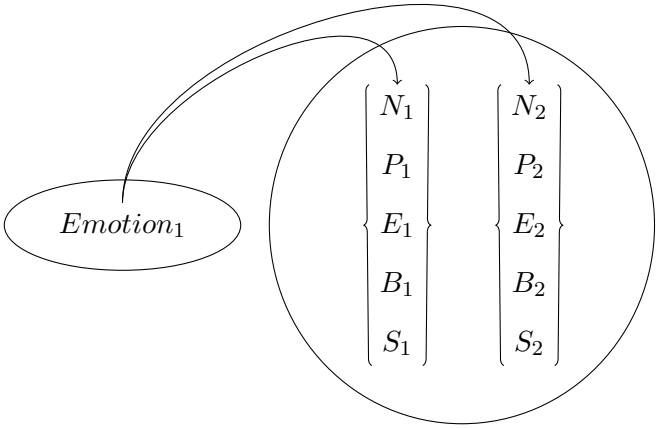
> *No one-to-one correspondence (NOC) thesis:* There is no one-to-one correspondence between anger, fear, happiness, sadness, and so forth, and any neurobiological, physiological, expressive, behavioral, or phenomenological responses.
>
> *Low coordination (LC) thesis:* There is low coordination between neurobiological, physiological, expressive, behavioral, or phenomenological responses among instances of anger, fear, happiness, sadness, and so forth. (Scarantino, 2015, p. 343)

On this presentation, NOC claims that emotions do not map one-to-one onto processes in the brain or the body, as well as expressions, behavior, or phenomenology. There are two ways in which this can occur. These are displayed in Figure 2.1. The first is a case (case (a)) where two emotion categories correspond to the same set of neural, physiological, expressive, behavioral, or phenomenological responses. The second (case (b)) is one where one emotion category is associated with two different sets of responses.

The LC thesis, in turn, holds that emotions do not constitute homogeneous processes in the sense that whatever neural, physiological, or behavioral underpinnings they may have, these are not coordinated in such a way as to allow their categorization under a single emotion category. In other words, the LC thesis states that there are no packages of responses at these levels that we can map onto emotion categories. In an earlier presentation of this claim, Scarantino and Griffiths (2011) explain:



(a)



(b)

Figure 2.1. Variations of NOC
(a) displays a case where two emotions correspond to the same set of neural (N), physiological (P), expressive (E), behavioral (B), and phenomenological (S) responses. (b) displays a case where one emotion corresponds to two different sets of responses.

> Evidence for LC consists of examples of anger, happiness, sadness, surprise, etcetera, that are instantiated in the absence of a coordinated package of physiological, neurobiological, expressive, behavioral, cognitive, and experiential responses. (Scarantino & Griffiths, 2011, p. 448).

This formulation is problematically ambiguous, as I will argue in detail below. In my view, the conditions under which a coordinated package of responses is absent are unclear. This makes it difficult to assess in which cases does LC holds, and what kind of empirical evidence would falsify it.

In the next section, I will discuss this problem as well as other conceptual issues with NOC and LC. In my view, these issues are not mere conceptual nuisances, but are paramount to evaluating whether VT is true. In turn, whether VT is true or not is vital to assess the status of emotion research and the prospects for a scientific theory of emotions. Thus, VT deserves special attention. Consequently, I will examine these

issues in detail and raise some questions that could help resolve them. This will help prepare the ground for what will come later in the chapter, where I will discuss in detail what I take to be the main problem with VT.

Before I move on to discussing these issues with VT, let us apply these theses to propose a working definition of the Empirical Challenge. As I explained before, the Empirical Challenge claims that a theory of emotions must be prepared to enable the rejection of VT. Given that VT is presented in terms of NOC and LC, we can expand our definition as follows:

Empirical Challenge Provide a scientifically meaningful theoretical framework that establishes correspondences between emotion categories and well-coordinated patterns of neural, physiological, expressive, behavioral, and phenomenological responses.

This formulation respects the empirical load VT is supposed to carry. VT is an empirical claim, thus requiring empirical means to reject it. If a proposed theory of emotions is successful in leading scientists to find patterns that correspond to emotion categories, we can say that the theory shows how emotions form unified phenomena and hence provides a framework to study emotions scientifically. Nevertheless, VT also requires conceptual work. It calls for an account of which types of responses are required to confirm or reject the claim.

With this working definition in hand, let us then move on to conceptual problems associated with VT. At the end of the chapter, I will revisit this working definition and apply the refinements I will propose to VT to offer a new Empirical Challenge. This will provide a standpoint from which to examine the prospects of overcoming or, as I will argue in Part II, dissolving this challenge.

2.1.1 Clarifying VT

VT, NOC, and LC

Let us start with some general problems regarding VT. As I presented it above, VT is constituted by two theses, NOC and LC. From the outset, this formulation already raises an important question: what is the logical relation between VT, NOC, and LC? To answer this question, one option is to look at Scarantino's formulations (Scarantino, 2015; Scarantino & Griffiths, 2011). However, these formulations are not helpful, since they limit themselves to claim that critics of basic emotion theories have proposed these two theses and that they are empirically supported. As a result, we must look elsewhere.

Regarding VT and NOC and LC, there are two initial possibilities: either VT is a conjunction or a disjunction of NOC and LC. Interpreting VT as a conjunction forces us to accept VT only in case there is no correspondence and no coordination. Yet, it seems that if there is no correspondence (NOC), this suffices to establish the claim that emotions are disjoined phenomena, even if there was coordination between different sets of responses. In other words, in case emotions map onto different coordinated

packages (NOC and not-LC), this still suffices to establish VT. This suggests that the correct interpretation is the disjunctive interpretation, rather than the conjunctive one.

However, in order to establish the disjunctive interpretation, we must also show that LC by itself is enough to imply VT. One way of approaching this issue is thinking of a case were LC is true and NOC is false. It is difficult to see such a case though. To put it clearly, this would require a case where there are no coordinated packages of responses associated with emotion categories (LC), and yet there is a one-to-one correspondence between the former and the latter (not-NOC). As it is clear from this formulation, this is contradictory. Hence, it seems that LC implies NOC. If there are no coordinated packages of responses, there is nothing to which emotion categories can correspond, making NOC trivially true.

Could there also be an implication relation from NOC to LC? It is easy to see why this is not the case. The fact that there is no correspondence between a given emotion category and a specific set of responses does not entail that there are no such coordinated sets. As in Figure 2.1b, there can be well defined coordinated packages of responses and yet no correspondence between them and emotion categories. Hence, NOC cannot imply LC.

So far LC implies NOC, and NOC does not imply LC. Yet, there is another potential relation between LC and NOC that deserves attention. Not only does the absence of coordinated packages (LC) imply the absence of correspondences (NOC), but only if there are such coordinated packages can correspondence be evaluated. In other words, it is plausible to think that correspondence only makes sense if there are packages to which emotions could potentially correspond. As a result, it would be necessary for LC to be false in order for NOC to be potentially true or false. This suggests that not-LC is a *presupposition* of NOC in Strawson's (1950; 2011) sense.[22] It must be true that there are coordinated packages of responses (i.e., false that there are no coordinated packages, not-LC) for correspondences to obtain (not-NOC) or not obtain (NOC). Notice that this analysis in terms of presuppositions clashes with our previous claim that LC implies NOC. This is because if NOC presupposes not-LC, NOC cannot have a truth value in case LC is true. Hence, LC cannot imply NOC.

How can we resolve the tension between these interpretations? This depends on how we interpret NOC. On a strong construal, NOC says that there are coordinated packages of responses but there is no correspondence between them and emotion categories. This leads to the presupposition interpretation, since the first of these conjuncts is not-LC. This has the additional consequence that NOC & LC is contradictory (since NOC requires the existence of coordinated packages which LC denies), and as a result leads to the reading VT as a disjunction: (NOC $\vee$ LC) $\equiv$ VT.

On a weaker construal, NOC is neutral regarding the existence of coordinated packages of responses. It merely states that there are no correspondences between

[22] Strawson's account of presupposition can be spelled out in the following terms, following van Fraassen (1968): "*A presupposes B* if and only if *A* is neither true nor false unless *B* is true" (van Fraassen, 1968, p. 137).

emotions and coordinated packages, but remains agnostic as to whether this is because there are no coordinated packages or whether there are such packages but they fail to correspond to emotions. On this reading, NOC does not presuppose not-LC, but LC does imply NOC, since the lack of coordinated packages is sufficient to establish lack of correspondence. Furthermore, it is still true that VT is true either because there are no coordinated packages (LC), which also implies that there is no correspondence, or because there are coordinated packages which fail to correspond. Thus, (NOC $\vee$ LC) $\equiv$ VT still holds on this reading.[23]

For the purposes of evaluating empirical evidence, we need not decide between either of these interpretations. As I have shown above, either of these interpretations leads to the disjunctive reading of VT. Furthermore, the only difference between these interpretations lies in whether we take NOC not to have a truth-value (strong reading) or to be true (weak reading) in case there are no coordinated packages of responses, a decision that does not have much bearing on empirical findings. Since either of these cases would support VT, I will leave the decision between these interpretations aside and proceed to examine NOC and LC independently.

NOC

Regarding NOC, let us note first that NOC admits a subdivision in terms in terms of the type of responses that emotions could correspond to, i.e., neural, physiological, expressive, behavioral, or phenomenological packages. Each of these NOC theses would claim that a given emotion does not correspond one-to-one with a given type of response (e.g. one emotion corresponding to two types of neural response). Thus, we can divide NOC into:

$\text{NOC}_{\text{Neural}}$ There is no one-to-one correspondence between emotion categories and any pattern of neurobiological responses.

$\text{NOC}_{\text{Physiological}}$ There is no one-to-one correspondence between emotion categories and any pattern of physiological responses.

$\text{NOC}_{\text{Behavioral}}$ There is no one-to-one correspondence between emotion categories and any pattern of behavioral responses.

$\text{NOC}_{\text{Expressive}}$ There is no one-to-one correspondence between emotion categories and any pattern of expressive responses.

$\text{NOC}_{\text{Phenomenological}}$ There is no one-to-one correspondence between emotion categories and any pattern of phenomenological responses.

By dividing NOC into these subtheses, we can have a better idea of what sources of empirical evidence would be relevant to test VT. Presumably, if an emotion category does not correspond to a single pattern of neural responses, for example, we have good reason to accept NOC.

[23] I thank Fabio Fang for helping me clarify this analysis.

However, this immediately raises in important issue: is NOC, as a general claim, a conjunction or a disjunction of the subtheses involved? Let (C) and (D) be these interpretations, respectively:

(C) $\text{NOC}_{\text{Neural}} \wedge \text{NOC}_{\text{Physiological}} \wedge \text{NOC}_{\text{Behavioral}} \wedge \text{NOC}_{\text{Expressive}} \wedge \text{NOC}_{\text{Phenomenological}}$

(D) $\text{NOC}_{\text{Neural}} \vee \text{NOC}_{\text{Physiological}} \vee \text{NOC}_{\text{Behavioral}} \vee \text{NOC}_{\text{Expressive}} \vee \text{NOC}_{\text{Phenomenological}}$

Even though I believe NOC is neither a conjunction or a disjunction of these claims, it is worth considering these options to understand what exactly is the correct interpretation of the claim. Interpreting NOC as a conjunction (C) leads to an overly simplified claim. As soon as we find correspondence in one domain, NOC will be false. For example, if an emotion fails to correspond to neural, physiological, expressive, and phenomenological sets of responses, but corresponds to one common behavioral pattern, NOC is falsified. Consequently, we would have to reject VT, given that there is at least one domain where correspondence holds. This is a consequence that defenders of VT would find unacceptable, given that an important degree of variation would still hold.

On the other hand, interpreting NOC as a disjunction of the different subtheses leads to an overly demanding claim to reject. In this case, evidence for lack of correspondence in one domain suffices to establish NOC and therefore to accept VT. Given that the aforementioned characterization of NOC includes domains where variability is expected (for example, in terms of action tendencies or expressions), rejecting NOC becomes not only implausible, but trivial. Evidence for some degree of variability in some domain abounds, rendering the question of variability almost insignificant.

In a more charitable interpretation, the various NOC subtheses have different weights. In other words, depending on the theory, some domains matter more than others. For instance, for traditional BET, lack of correspondence in the neural or physiological domains ($\text{NOC}_{\text{Neural}}$ and $\text{NOC}_{\text{Physiological}}$) suffice to claim that an emotion does not correspond to a coordinated package of (the relevant) responses, where as lack of correspondence in the behavioral outcomes $\text{NOC}_{\text{Behavioral}}$ by itself would not suffice. Discrete appraisal theorists, on the other hand, may accept lack of correspondence in the first domains without accepting VT, but may have difficulties doing so with the latter domain.

In order to escape these problems, researchers must decide which domains offer the most relevant support for NOC. Such consensus, however, is lacking in the literature. This makes it difficult to assess NOC and hence VT in order to decide between different theories of emotions, since each of these theories have different decision criteria. In the later sections, I will offer arguments to limit NOC to the neural, physiological, and behavioral domains, and suggest some ways in which we can reach a neutral account of variability that helps us decide between different theories. Before this, however, let us turn to LC.

LC

As I explained above, LC is formulated in terms of the absence of a coordinated package of responses. What is it for there to be such an absence? On a perhaps overly literal reading of this claim, this would be a case where an emotion obtains but no responses are observed. But then, why would we accept that an emotion obtains? If there are no neurological, physiological, behavioral, expressive, or phenomenological responses, there is no emotion either. It seems clear that this is not the intended interpretation. Yet, it is the reading that stems from a literal reading of Scarantino's and Griffiths's formulation.

If we do not read this claim as stating that there is no set of responses at all, what does it mean for a coordinated package of responses to be absent? In its most plausible interpretation, LC is a claim about correlations. On this reading, a coordinated package of responses is absent in case the different associated are not robustly correlated with one another. Barrett (2006), presenting some evidence for this claim, writes:

> Although no single study of emotion has simultaneously measured facial movements, vocal signals, changes in peripheral physiology, voluntary action, and subjective experience, many studies have measured at least two or three of these responses (usually some combination of subjective experience, behavior, and autonomic activity). These studies have reported a range of associations, from modest correlations to no relationship to negative correlations among experiential, behavioral, and physiological measures of emotion. (Barrett, 2006, p. 33)

Following this interpretation, LC is a thesis about the correlation between different measurements. LC would be thus established if for a given emotion category, we fail to find that, for instance, skin conductance responses (physiological measure) for anger do not correlate with anger expressions, or that neural activity for sadness fails to correlate with retreat action tendencies characteristic of the emotion (behavioral measure).

Even though this is a step forward towards a scientifically tractable interpretation of the claim, it needs further requirement. First, given its appeal to correlations, it is unclear which correlations (or lack thereof) are necessary or sufficient to reject (or accept) LC. There are two senses in which we can raise questions about correlations between measurements. In one sense, we can ask about correlations between variables in the same domain. Call this *intra-domain coordination*. In another sense, we can ask about correlations between variables of different domains. Call this *inter-domain coordination*.

Intra-domain coordination would obtain in case variables of the same domain are robustly correlated with one another. However, even inside a single domain, there is a myriad of possible variables to measure. Consider the physiological domain. Among physiological measures used to study emotions, we find three families (cardiovascular, respiratory, and electrodermal), each with a wide range of possible measurements. Given the different possible variables researchers could employ in their studies, we can

ask: do we require correlations between all of these variables in all of these domains in order to reject LC? If not, which correlations suffice? And to which degree?

Additionally, many of these variables will be causally related to one another. For example, increased heart rate will cause increased respiratory frequencies. This makes certain correlations trivial, as two causally related variables will always be robustly correlated. While these causal relations are interesting to individuate (in this case physiological) patterns of responses, this forces a reformulation of the coordination thesis. On one hand, there are coordinated packages of responses in a given domain (i.e., there is intra-domain coordination) in case we have a set of causal relations that are well distinguished from other sets (thus forming a package or a pattern), and, in case there are no strict causal relations between two variables, we observe a robust degree of correlation.

Concerning inter-domain coordination, similar problems arise. We can expect causal relations between variables in different domains. For example, we may expect certain expressive patterns such as blushing to be correlated with physiological variables related to cardiac processes, since blushing is increased capillary blood flow in the skin in certain parts of the face. Hence, we must also specify in which cases does inter-domain coordination obtains in terms of both sets of causal relations and, in case two sets of variables in different domains are not causally related, a robust degree of correlation.

There is an additional problem regarding inter-domain coordination that must be addressed. Suppose that a given neural pattern, say amygdala activity, is robustly correlated with different behavioral responses, e.g. fight, flight, or fleeing responses (a candidate construal of fear reactions). Should we conclude that there is low inter-domain coordination because one neural pattern is associated with three distinct behavioral responses? If so, then inter-domain coordination is almost impossible, since there are domains in which variations are to be expected, such as the behavioral and expressive domains (relative to the neural and physiological domains). As a result, LC becomes almost trivial, since it is easy to find cases where inter-domain coordination does not obtain on this criterion.

To solve these problems, we must be clear on two fronts. First, we must have clear criteria of robustness. This is to say, we need to have clear conditions under which we will say that coordination (intra-domain or inter-domain) fails to obtain. This depends on how we interpret correlations, at the very least, which would require a detailed discussion that I will not pursue here. For now, we can assume that such a criterion can be given by a proper methodological analysis.

Assuming such a criterion, the second front we must be clear about regards the individuation of patterns in a given domain. In other words, we must have an idea of what counts as a neural, physiological, expressive, behavioral, and phenomenological pattern. This is because questions about whether or not there is inter-domain variability depend on whether we consider different variations parts of a common pattern or not. In other words, this is a question about how coarse or fine-grained are we going to cut our patterns. In the example above, whether or not fight, flight, or freezing re-

actions are considered as instantiating a common behavioral pattern such as "avoiding danger" will help decide whether one neural pattern is said to be associated with one or three patterns of response.

This problem of pattern individuation does not only affect the case for inter-domain coordination, but also plays and important role in inter-domain coordination and questions about correspondence. In order to decide whether or not a set of causal relations and correlations counts as one pattern (intra-domain coordination) and to what kinds of patterns should emotions correspond to evaluate NOC, we must have an account of pattern individuation for each domain.

In what follows, I will explore the problem of pattern individuation by adopting a naturalistic, empirically informed perspective. I will examine how scientists have attempted to individuate patterns in each domain. In some cases, I will offer arguments for a specific criterion of pattern individuation. In others, I will suggest ways of offering such criteria, although a final decision would require extensive work and further investigation.

In any case, this approach will also help us establish whether or not NOC and LC are empirically well supported. The discussion above offers us a working definition of these theses already. Regarding NOC, I shall leave it as formulated by Scarantino, noting that correspondences will depend on the existence of coordinated patterns of responses according to the criteria corresponding to each domain, and to relevance criteria regarding which domains are relevant to distinguishing emotions from one another. With regards to LC, however, a reformulation is in order. I propose the following working definitions of these claims:

No one-to-one correspondence thesis (NOC) There is no one-to-one correspondence between emotion categories and neural, physiological, behavioral, expressive, or phenomenological patterns of responses.

Low coordination thesis (LC*) Variables in the neural, physiological, behavioral, expressive, and phenomenological domains do not constitute well-defined patterns of responses (i.e., display low correlations and do not constitute a well-defined set of causal relations either among variables in a domain or between variables in different domains).

2.2 Individuating response patterns

2.2.1 Neural patterns

Traditional accounts of emotion that emphasized the role of neural mechanisms in emotions thought of emotions as relating to the activity in specific and consistent regions in the brain. According to these views, there must be something in the brain that is domain-specific to each emotion category. These are for example LeDoux's studies on fear conditioning (LeDoux, 2000, 2003, 2007, 2013, see also Phelps and LeDoux 2005),

which attempted to map fear onto amygdala activity, or Panksepp's (1998; 2011) attempt to individuate subcortical structures underlying primary emotional processes. Following Lindquist, Wager, Kober, Bliss-Moreau, and Barrett (2012), let's call these locationist accounts:

> [Locationist accounts] hypothesize that all mental states belonging to the same emotion category (e.g., fear) are produced by activity that is consistently and specifically associated with an architecturally defined brain locale [...] or anatomically defined networks of locales that are inherited and shared with other mammalian species. (Lindquist et al., 2012, pp. 122-123)

From this formulation of locationism, we can distinguish two strands:

Anatomical locationism Mental states belonging to the same emotion category correspond consistently and specifically to an *architecturally defined* brain region.

Homological locationism Mental states belonging to the same emotion category correspond consistently and specifically to *inherited networks in the brain shared with other mammalian species.*

Anatomical locationism

Anatomical locationism is arguably the most traditional form of locationism. It attempts to map emotions onto distinct brain regions, individuated in virtue of their architecture. This position is, for many, the most intuitive form of locationism. As a result, the first meta-analyses on emotions in the brain attempt to test whether there is good evidence for such a mapping.

One of the earlier, if not the first, meta-analysis in this regard was the one by Phan, Wager, Taylor, and Liberzon (2002). This meta-analysis examined findings across imaging studies in search of specific regions associated with emotional activation in general, with specific emotions and different induction methods. In particular, they examined how sensitive specific brain regions were to different emotional tasks and how specific certain brain regions were for different emotional responses. The researchers took PET and fMRI studies from January 1990 to December 2000, out of which 55 studies from May 1993 to December 2000 met the criteria.

First, they found that no specific brain region was consistently activated in the majority of studies, across individual emotions and induction methods. This implies that no single brain region is commonly activated by all emotional tasks. Specifically, no particular region was activated in over 50% of the tasks. Nevertheless, the mPFC was often activated, albeit not specific to a given emotion or induction method. The authors take this to reflect that certain aspects may be shared across multiple individual emotions. This is supported by the finding that the mPFC is involved in 4 out of 5 emotions tested (happiness, anger, sadness, and disgust, but not in fear) in at least 40% of studies. Given this result, the researchers argue that this region may

be involved in cognitive aspects of emotion processing, such as attention to emotion, appraisal, or emotion identification.

Concerning particular emotions, they report some interesting results for fear, sadness, happiness, and disgust. For fear, the researchers found a strong association with amygdala activity (i.e., reported activity in around 60% of studies). In these studies, the amygdala showed activation for the recognition of fearful facial expressions, feelings of fear after procaine induction, fear conditioning, evocation of fearful emotional responses from direct stimulation, detection of environment threat, and the coordination of appropriate responses to threat and danger. A similar pattern follows for sadness, which was associated with activity in the subcallosal cingulate cortex (46% of the studies). On the other hand, results were not so positive for happiness and disgust, which were both linked to activation in the basal ganglia (70% of studies and 60% of the studies, respectively).

Despite some initial promising results, Phan et al.'s results also point towards some lack of specificity. For example, the fact that both happiness and disgust show basal ganglia activation suggests that these areas are not unique to either emotion, but rather are involved in some general process. Similarly, the researchers discuss the association between fear and the amygdala, hypothesizing that this region might play more general role involving vigilance or processing salience. This is supported by the fact that the researchers have found amygdala activation for all kinds of evocative stimuli, including fear faces, aversive pictures, sad and happy faces, and positive pictures.

Hence, it would appear as if empirical evidence supported some degree of intra-domain coordination in the neural case but only for a few emotions (given the aforementioned mappings regarding fear, sadness, happiness and disgust). Consequently, as far as this meta-analysis is concerned, correspondences also fail to obtain overall, in spite of some correspondences being found.

At around the same time as Phan et al., Murphy, Nimmo-Smith, and Lawrence (2003) also published a meta-analysis studying the neural specificity of emotion. In particular, Murphy and colleagues investigated whether there are differences in three-dimensional distributions of neural activity associated with (1) positively vs. negatively valenced emotions, (2) approach vs. withdrawal emotions, and (3) affect program emotions (fear, anger, disgust, happiness, and sadness), and whether there are specific associations between certain affect program emotions (fear and disgust) and specific neural regions (amygdala and insula/basal ganglia), among other hypotheses. They included 106 PET or fMRI studies ranging from January 1994 to December 2001.

They found that distributions between negative and positive emotions did not differ significantly in any condition, and that there was a significant difference between approach and withdrawal emotions (greater left- than right-sided activation for the former). Concerning the distinctions between what they called affect program emotions, they report that all emotions differed from each other significantly based on their spatial distribution, except for happiness which only approached significance against anger and did not differ significantly from sadness. As for their location, they report that the most consistently activated regions for each emotion were the amygdala for

fear, the insula/operculum and the globus pallidus for disgust, the lateral OFC for anger, the rostral supracallosal cortex for happiness, and the rostral supracallosal but also the anterior cingulate cortex and the dorsomedial PFC for sadness.

Building on these two previous meta-analyses, Vytal and Hamann (2010) made another meta-analysis updating and sophisticating the methods of the previous ones. The researchers conducted two types of analyses: consistency and discriminability analyses. Consistency analysis determined the brain regions whose activity was mostly consistently and strongly associated with each of the individual basic emotions. Discriminability analysis contrasted activations associated with each basic emotions to assess whether patterns of regional brain activation can discriminate between different basic emotions. They investigated neuroimaging studies that included either explicit emotional elicitation tasks (e.g. mood induction), emotionally arousing stimuli (e.g. emotional pictures), or emotional facial expressions. They selected 83 PET or fMRI studies from 1993 to 2008.

Consistency analysis revealed a number of clusters for each emotion. For happiness they found 9 clusters, with the largest being located in the right superior temporal gyrus; for sadness, 35 clusters, the largest in the left medial frontal gyrus; for anger, 13 clusters, the largest in the left inferior frontal gyrus; for fear, 11 clusters, largest in the amygdala; and for disgust, 16 clusters, largest in the right insula and right inferior frontal gyrus. As for their discriminability analysis, the researchers also observed a number of clusters that distinguish each emotion from the others significantly. The locations of these clusters are summarized in Table 2.1. Given that there are significant clusters of activation in specific locations in the brain, the authors claim that this meta-analysis provides evidence for the neural specificity of the aforementioned emotions.

Furthermore, Vytal and Hamann ran the same analyses using the same dataset as Murphy et al. (2003), and found that not only were the previous findings replicable, but they could also differentiate between happiness and sadness, a distinction that the previous meta-analysis did not find. Consequently, Vytal and Hamann conclude that given more sophisticated methods and analytic tools, we can find specific regions in the brain whose activity corresponds to and helps distinguish between different emotions. As for the mapping of emotions onto brain regions, they suggest the following mapping:

Happiness Rostral ACC and right STG

Sadness MFG and head of the caudate/subgenual ACC

Anger IFG and Parahippocampal gyrus

Fear Amygdala and insula

Disgust IFG/Anterior insula

Vytal and Hamman's results suggested a more optimistic outlook for coordination and correspondence from the perspective of anatomical locationism. Now there was a set of emotions which were mapped to more or less distinct brain regions. Even though

	Happiness	Sadness	Anger	Fear	Disgust
Happiness	–	R. STG	L. rACC	R. STG	L. rACC
Sadness	R. MTG	–	L. MFG	L. mFG	R. IFG
Anger	IFG	R. parahippocampal gyrus	–	L. IFG	L. IFG
Fear	L. amygdala	L. amygdala	L. putamen	–	L. amygdala
Disgust	R. putamen	L. insula	R. putamen	R. putamen	–

Table 2.1

Summary of results of the discriminability analysis by Vytal and Hamann (2010). Only the location of the largest cluster found is reproduced here. The following abbreviations are also used: L = Left, R = Right, r = rostral, m = medial, STG = superior temporal gyrus, ACC = anterior cingulate cortex, MTG = middle temporal gyrus, MFG = middle frontal gyrus, IFG = inferior frontal gyrus.

some emotions were associated with two different regions, this could still be interpreted as an invitation for further research to discover which region was responsible for the given emotion.

In spite of this optimism, it was not long before other researchers presented results in the opposite direction. Especially on the constructionist side of the debate, neural specificity fell under attack. First, Barrett's (2006) review criticized previous studies and meta-analyses, arguing that they did not support the case for specificity. Later, another meta-analysis was published, targeting these meta-analyses in detail. This was Lindquist et al. (2012).

Lindquist et al.'s meta-analyses set out to examine evidence for and against locationist views, contrasting them with constructionism. They characterize the locationist view as committed to the claim that "instances of an emotion category (e.g. *fear*) are *consistently* and *specifically* associated with increased activity in a brain region (or a set of regions within an anatomically inspired network)" (Lindquist et al., 2012, p. 126).[24] By consistency, they mean "the fact that a brain region shows increased activity for every instance of an emotion category" (ibid.); by specificity, "the fact that a given brain region is active for instances of one (and only one) emotion category" (ibid.). In contrast to locationism, constructionism is taken to reject both of these hypotheses, instead claiming that "the same brain region(s) are more generally important to realizing a basic psychological operation (e.g. core affect, conceptualization, language, or executive attention)" (ibid.).

[24] This definition includes both versions of locationism presented above. Nevertheless, I will only consider evidence against anatomical locationism at the moment.

In their meta-analysis, Lindquist and colleagues begin by estimating the neural reference space assuming a discrete classification of emotions. This neural reference space is comprised of brain regions that show a consistent increase in activation for instances of anger, sadness, fear, disgust, and happiness, or for the entire category of emotion. Using this neural reference space, they examined each region's consistency with each emotion category (e.g. whether the amygdala is consistently associated with all instances of fear). Additionally, they asked whether there was an absolute difference in the proportion of contrasts activating near those voxels for each emotion category versus the others. This provided grounds to test the specificity claim, that is, whether the voxels that are activated for all instances of an emotion category (consistency) are also active selectively for that category. The researchers present their results in a region to region basis.

They start off with the amygdala, a region traditionally associated with fear, as seen in the meta-analyses presented above. Lindquist et al. argue that rather than being associated with fear, the amygdala is part of a network that helps realize core affect.[25] More precisely, they claim that the amygdala is involved in "signaling whether exteroceptive sensory information is motivationally salient" (Lindquist et al., 2012, p. 130). The reason for this is that amygdala activity is also observed in other tasks such as orienting responses to motivationally relevant stimuli, novel and unusual stimuli, and that lesions to the amygdala do not only affect fear responses, but also responses to other relevant stimuli in general. Additionally, their analyses reveal that amygdala activity is also significantly associated with disgust.

Next, Lindquist and colleagues discuss the anterior insula, previously associated with disgust. According to the constructionist hypothesis, the anterior insula is involved in "representing core affective feelings in awareness [i.e.] awareness of bodily sensations [...] and affective feelings" (Lindquist et al., 2012, p. 133). In this line of argument, they claim that the anterior insula shows increased activation during awareness of bodily movement, gastric distention, and orgasm. Electrical stimulation to the anterior insula also produces other feelings besides disgust, such as feelings of movement, twitching, warmth and tingling in the lips, tongue, teeth, arms, hands, and fingers. The reseachers also add that the left anterior insula shows increased activation in instances of anger.

Regarding the OFC, previously associated with anger, Lindquist et al. claim that we should understand its role as the integration of exteroceptive and interoceptive sensory information to guide behavior. Evidence for this claim comes from studies showing that the lOFC and the mOFC have been linked to associative learning, decision making, and reversal learning. Additionally, against the claim that the OFC shows specificity for anger, the left OFC shows increased activation in instances of disgust.

[25] Recall that core affect is the "neurophysiological state consciously accessible as the simplest raw (nonreflective) feelings evident in moods and emotions" (Russell, 2003, p. 148). See also §1.2.3, p. 37.

Lastly, the authors consider the ACC, which was thought to be associated with sadness. In contrast, Lindquist et al. hold that the ACC is involved in a more general process of visceral regulation. According to their meta-analysis, the ACC displays increased activity in a number of instances including mania (in contrast to depression), executive attention and motor engagement. All of these processes involve some kind of regulation of somatosensory states, which would support the constructionists hypothesis.

Besides the regions previously linked to basic emotions, psychological constructionists expect a number of other regions to be involved in emotion processing. These other regions would be involved in conceptualization, memory, and other processes related to psychological construction. Among the evidence they cite for these regions we find the dmPFC, MTL, and retrosplenial cortex as being involved in a conceptualization network. This would provide evidence that there is no system nor a set of regions that are specific and consistent either to particular emotion categories nor to emotion in general. Rather, there are a variety of systems that are active depending on different more general domain processes, constructionists claim.

Lindquist et al.'s influential meta-analysis showed that on the anatomical locationist framework, coordination and correspondence could not be established for the neural domain. In my view, their evidence is sound, although I do not endorse their conclusion. For Lindquist and colleagues, evidence supports a constructionist theory of emotions which assumes that VT is already well-established. I contend that further argument is required to establish VT though. On one hand, evidence regarding homological locationism is still required, since it may be the case that neural patterns corresponding to emotions are at the network level, rather than at the region level. Let us explore evidence in this direction.

Homological locationism

As explained above, homological locationism attempts to map emotion categories onto intrinsic networks, rather than specific brain regions. To understand homological locationism, it is worth attending to the distinction between intrinsic and functional connectivity. A network is said to be intrinsic in case it is hardwired in the brain, that is, if it is a set of brain regions that work together because of predetermined pathways of excitation and inhibition. This is in contrast to functional networks, which are correlated but not necessarily in virtue of predetermined mechanisms.

One example of a homological locationist view is Panksepp's (1998; 2011). Panksepp claims that in order to individuate the neural patterns in the brain, we must rely on comparative studies using non-human animals to find evolutionarily adapted networks. These comparative studies would presumably lead to the discovery of intrinsic networks shared with other species, which Panksepp calls *emotive circuits*. These circuits are defined as follows:

1. The underlying circuits are genetically predetermined and designed to respond unconditionally to stimuli arising from major life-challenging circumstances.
2. These circuits organize diverse behaviors by activating or inhibiting motor subroutines and concurrent autonomic-hormonal changes that have proved adaptive in the face of such life-challenging circumstances during the evolutionary history of the species.
3. Emotive circuits change the sensitivities of sensory systems that are relevant for the behavioral sequences that have been aroused.
4. Neural activity of emotive circuits outlasts the precipitating circumstances.
5. Emotive circuits can come under the conditional control of emotionally neutral environmental stimuli.
6. Emotive circuits have reciprocal interactions with the brain mechanisms that elaborate higher decision-making processes and consciousness.
   (Panksepp, 1998, pp. 48-49)

On this definition, emotion circuits are not meant in terms of specific regions, but rather collections of regions that work connectedly. What is important is that these networks are "genetically predetermined," meaning that they depend on intrinsic properties of the brain rather than functionally constructed networks (which I will discuss below).

Panksepp (2011) presents evidence for subcortical networks that interconnect midbrain circuits with various structures in the basal ganglia, such as the amygdala and the nucleus accumbens, through pathways running through the hypothalamus and thalamus. He characterizes each of these networks by using capitalized labels to remain agnostic about their correspondence with vernacular emotion terms. He writes:

> These labels, by using full-capitalization of terms, refer to specific subcortical networks in mammalian brains that promote specific categories of built-in emotional actions and associated feelings. No claim is made of identity with the corresponding vernacular words, although profound homologies are anticipated. (Panksepp, 2011, p. 8)

Among the networks he identifies, he includes:

SEEKING Nucleus accumbens – VTA, Mesolimbic and mesocortical outputs, Lateral hypothalamus – PAG.

RAGE Medial amygdala to Bed Nucleus of Stria Terminalis (BNST), Medial and preformical hypothalamic to PAG.

FEAR Central and lateral amygdala to medial hypothalamus and dorsal PAG.

LUST Cortico-medial amygdala, Bed nucleus of stria terminalis (BNST), Preoptic hypothalamus, VMH, PAG.

CARE Anterior cingulate, BNST, Preoptic Area, VTA, PAG.

PANIC Anterior cingulate, BNST and Preoptic Area, Dorsomedial Thalamus, PAG

JOY Dorso-medial diencephalon, Parafascicular Area, PAG. (adapted from Panksepp, 2011, p. 9)

In spite of Panksepp's optimisim, there is nevertheless evidence against the presence of intrinsic networks in the brain as well. Touroutoglou, Lindquist, Dickerson, and Barrett (2014) show that increases in activity during emotion experience and perception do not map onto intrinsic networks in the brain using resting state connectivity fMRI. Their meta-analysis is based on the previous work by Vytal and Hamann (2010) mentioned above. Touroutoglou and colleagues write:

> If anatomically constrained networks for each emotion category exist in the intrinsic architecture of the human brain, as the basic emotion view [e.g. Panksepp's] predicts, then the meta-analytically derived seed regions for a given emotion category (i.e. the peaks of consistent activation for a given category of emotion, such as happiness) should produce 'discovery' maps whose spatial overlap reveals a network for that category. This finding would provide strong support for the hypothesis that emotions are biologically basic categories reflected in the intrinsic structure of the brain. Alternatively, if the peaks observed in Vytal and Hamannn (2010) are nodes in domain-general intrinsic networks, as predicted by the conceptual act theory of emotion [now called the Theory of Constructed Emotion], then the conjunction of the discovery maps for a given emotion category would not converge on a single network. Instead, emotion-based seeds would give evidence of the domain-general intrinsic networks that are already known to exist in the literature. (Touroutoglou et al., 2014, p. 1258)

The idea then is that if there are intrinsic networks individuating each emotion category, the activation maps for each emotion category should consistently reveal a specific set of regions that are intrinsically connected. The conjunction of sets of active regions for a given emotion category is what they call the *discovery map*. If coordination in terms of intrinsic networks fails to obtain, we should expect that for a given emotion category, there are a variety of networks which are presumably domain-general, hence yielding an empty discovery map.

The researchers report finding domain-general networks involved in emotional experience, which in their view supports the claim that there is nothing we can call a coordinated package of neural responses in terms of intrinsic networks. For example, for anger, they observed no overlap between the sets of active regions, suggesting an empty discovery map. For fear, sadness, and happiness, they found a general dorsal region connecting the anterior insula and the anterior cingulate cortex.

These results suggest that coordination in terms of intrinsic networks, i.e., homological locationism, is not empirically well-supported. In other words, we cannot obtain coordinated patterns of neural activity at the level of intrinsic networks. As with previous efforts to establish variability, the authors take this to support constructionism.

Yet again, there is still an alternative approach. As I explained above, there are two senses in which we can talk about brain networks. One is by identifying networks that are genetically predetermined or hardwired. A second approach is to think of networks in terms of their functional connectivity patterns. In this case, we can expect sets of active regions that correlate robustly with one another, even if their connections are not wired from birth. Call this strategy *pattern assignment*:

Pattern assignment Mental states belonging to the same emotion category correspond consistently and specifically to *functionally individuated patterns in the brain.*

Let us explore the prospects for such an approach.

Pattern assignment

As I just presented the view, instead of trying to individuate patterns in terms of domain-specific intrinsic networks, we could take the relevant neural patterns to be at the level of regions that show correlated activation even in the absence of an intrinsic network, e.g., functional networks or distributed patterns of activation. A prime example of this approach is the one involved in multivariate pattern analyses (MVPA).

Multivariate pattern analyses of brain activity steer away from traditional locationist hypotheses that try to map psychological states onto specific regions in the brain. Instead, these analyses try to find networks of brain activity that underlie psychological states. By focusing on the interaction between different regions, multivariate techniques allow researchers to identify higher level patterns of activation, thus overcoming failures in one-to-one mappings of psychological states and brain regions. In the case of emotions, multivariate techniques provide an interesting reply to the constructionist challenge. As I explained before, constructionists stress the lack of neural specificity for emotions in terms of brain regions. However, it is possible to accept this type of variability while expecting specificity at a network level.

One of the earlier studies using these techniques is Kassam, Markey, Cherkassky, Loewenstein, and Just (2013). Kassam and colleagues scanned ten subjects' brains using fMRI. During the scan, the subject saw different emotion words, and they were asked to attain the corresponding emotional state for a period of time. The researchers then trained a classifier on the data in order to test whether the classifier could accurately predict the subject's emotional state using only their neural activity data.

Kassam et al. report that their classifier was able to successfully predict a subject's emotional state from their neural data in a given trial. In other words, using a subject's data throughout the experiment, they could predict their emotional state in one trial

by looking at their neural activation. They report that this classification was accurate between 77% and 89% of the time. Besides this result, the researchers attempted to predict a subject's emotional state using all of the participant's data. This yielded a lower, but still significant accuracy score of 70%. Kassam et al. could even predict the emotional content of a stimulus from a different modality with their classifier. Specifically, they used data obtained with their word-cued induction method to see if they could predict the content of a visual image. As they report, the classifier was able to predict disgust 91% of the times. Even though this is only a significant finding for one emotion, they claim that their results were promising enough to support further research.

After Kassam et al. (2013), other studies followed suit. Kragel and LaBar (2015) tested how multi-voxel patterns of BOLD response predict discrete emotional states and whether these patterns conform to categorical or dimensional models of emotion. They used film emotion induction and instrumental music induction followed by self-report. The researchers classified seven emotional states (contentment, amusement, surprise, fear, anger, sadness, and neutral) with 37.3% accuracy (chance = 14.3%). These measures, they claim, show reliable detection of emotion-related information.

To test whether neural classification models generalized across modalities, they trained the classifier on the film data and tested on the music clips. This yielded an accuracy rating of 28.38% (again, chance = 14.3%). They then tested which voxels contributed to the models' accuracy the most, and report a variety of voxels throughout the brain. They explain that the patterns seem distinct yet partially overlapping at a macro-scale, as activation within many of the same structures contributed to the prediction. However, at the voxel level, the patterns did not overlap significantly.

To compare between categorical and dimensional models, they created a model for each type. The categorical model divided the self-report data into seven dimensions, corresponding to each discrete emotion. The dimensional model reduced the data to valence and arousal. They calculated how homogeneous the data was in each model. Accordingly, they report that categorical models yield more sparse and equidistant clusters, whereas the dimensional model yields a more clustered and overlapping view.

The researchers conclude that we can successfully predict the occurrence of seven distinct emotional states using MVPA across two induction methods. Regarding categorical vs. dimensional models, they conclude that it is best to see them as complementary, given that both models are able to classify emotions successfully in terms of self-report, although the dimensional model did not do well for the neural data. For the purposes of testing coordination in the neural domain, this study supports the pattern assignment strategy, since it is possible to obtain categorical classification in terms of multivariate patterns at the neural level.

Most recently, one study has gained special attention among defenders of multivariate approaches to emotion: Saarimäki et al. (2018). In this study, the researchers investigated the neural underpinnings of different basic and non-basic emotion categories using fMRI. Participants heard 4 narratives for each of 14 emotional states plus a neutral condition. Each narrative was designed to elicit the corresponding emo-

tion. Using MVPA, the researchers then trained a classifier for each participant and afterwards averaged across participants.

At the behavioral level, ratings showed that the narratives successfully elicited the target emotions reliably and strongly. Accuracy assigning a narrative to the correct target category was 97%. Regarding the classification of basic vs. non-basic emotions, the researchers report a mean classification accuracy across the 14 emotions and neutral state was 17% (chance level of 6.7%) and above significance when corrected for multiple comparisons. On average, basic emotions could be classified more accurately than non-basic emotions (26% vs. 15% respectively).

According to the researchers, experiential similarity matrices derived from behavioral ratings was significantly associated with the neural similarity matrices derived from whole-brain classification. Clustering of whole-brain confusion matrices yielded four clusters:

1. Happiness, pride, gratitude, love, and longing
2. Disgust, sadness, fear, and shame
3. Anger, contempt, guilt, and despair
4. Surprise and neutral

Positive emotions in cluster 1 were more prominent in the anterior frontal areas, including the vmPFC. Negative emotions in cluster 2, in turn, were more prominent in the insula, supplementary motor area, and specific parts of subcortical structures. Negative social emotions in cluster 3 were associated with the left insula and adjacent frontal areas, and surprise was associated with the auditory cortex, supplementary motor areas, and left insula.

Saarimäki et al. conclude that multiple emotion states have distinct and distributed neural bases. In their view, many emotions are represented in the brain in distinct yet overlapping regions. They claim that each emotion state likely modulates different patterns measured with fMRI, and the overall configuration of the regional activation patterns defines the resulting emotion. Altogether, 12 emotions (excluding longing and shame) could be reliably classified from fMRI signals. Yet, the researchers are careful to clarify that this study cannot provide causal evidence, given its use of classification techniques. These, they claim, require lesion studies rather than mere classification. Among the regions that gave rise to different emotions, the researchers report the ACC, PCC, and precuneus for most emotions. They also found activation in the brainstem, including periaqueductal grey, pons, and medulla, which they interpret as probably modulating autonomic responses.

These results from MVPA offer a lifeline for theories claiming neural specificity of emotion. In other words, they seem to support some form of coordination in the neural domain. Given the presence of networks mapping onto emotion categories, it is possible for theorists to cash out specificity at a network level, bypassing the demand for mappings at the level of brain regions. As McCafrey (in press) puts it regarding basic emotion theory:

> [...] the brain may respect BET after all. While individual regions are recruited for multiple emotions, [basic emotions] may map onto distributed, overlapping brain networks. We should therefore abandon the assumption that BET requires whole distinct neural circuitry for BEs.
>
> [...] multivariate analyses may provide valuable new tools for finding BE biomarkers. While individual variables (e.g. heart rate, face muscles, brain areas) are often non-specific for emotions, patterns among multiple variables tend to improve specificity. (pp. 29-30)

In spite of McCaffrey's and others' optimism, there are also reactions against these approaches. Clark-Polner, Johnson, and Barrett (2017), for instance, argue that MVPA findings do not support specificity at all. In their view, the patterns found using MVPA do not reveal brain states underlying specific processes, but rather a "statistical summary" of what goes on in the brain frequently under certain conditions. In their words:

> Any statistical summary of a category is an abstraction that does not necessarily exist in nature. This is also how emotion categories work [...]. Although as a group the instances of any emotion category can be diagnosed with a pattern, the pattern itself is an abstraction and does not necessarily describe one feature or set of features that is necessary for every (or even any) single individual instance in the category. [...] Thus, in a statistical sense, Saarimaki [sic] et al. did not find evidence to support the theory of basic emotion, nor the existence of biologically basic emotion categories. (Clark-Polner et al., 2017, p. 1946)

What does it mean for patterns to be a statistical summary? Clark-Polner and colleagues offer the following example. Consider the statement "the average middle class U.S. family has 3.13 children." This statement does not say that every middle class U.S. family has 3.13 children. Moreover, this interpretation would be non-sensical, as a family cannot have non-integer numbers of children. What this statement expresses is a statistical fact about middle class U.S. families, namely, the fact that the average number of children is 3.13. A similar logic, they claim, applies to patterns found with MVPA. When we find a set of voxels associated with a given process (in this case, a given emotion category), all we are saying is that in average, these voxels are active during the emotional episode. This does not mean that any of these voxels is necessary or plays a relevant role in the ongoing emotion.

It is true that statistical summaries do not imply any causal property of the object of study, as Clark-Polner and colleagues argue. The fact that the average middle class U.S. family has 3.13 children does not say anything specific about any particular family in the U.S. But it is not true that statistical summaries do not say anything interesting about their objects. In the example above, knowing this statistical fact leads us to infer that any given middle class U.S. family will most likely have a number of children

ranging from 0 to 6, approximately. Of course, statistical summaries allow exceptions, but there are still properties that we can infer from them, at least probabilistically.

A similar argument can run for MVPA: voxels constituting patterns associated with an emotion may not be present for every single instance, but the fact that they are still associated with it helps us map certain emotion categories to broadly construed brain networks. We need not commit ourselves to the claim that every single voxel in the pattern will appear in every single instance. We can do with the statistical summary and infer that there are typical mechanisms, specified in terms of networks, that do map onto particular emotion categories.

Variability at the neural level

Overall, it seems that the best candidate to spell out patterns of neural activity that would support coordination is in terms of pattern assignment. Evidence against anatomical and homological locationism seems compelling, or at the very least plausible. If researchers want to establish coordination and correspondence in the neural domain, characterizing neural patterns in terms of functional networks seems the most attractive remaining alternative.

However, the pattern assignment strategy is not without problems. As critics of functionalism such as Clark-Polner et al. point out, it might be the case that mapping psychological function, in this case emotion categories, onto functional networks is not explanatorily interesting. In my view, statistical summaries do constitute explanatorily relevant findings about how emotions are realized in the brain, and hence should be taken seriously. Nevertheless, a full-fledged defense of this claim requires a detailed discussion about how to map psychological function onto the brain.

This question, in its broad construal, is the question at the center of debates on *cognitive ontology*. Two questions in these debates are of vital interest for our current discussion. The first is the question raised above, namely, to which degree are functional networks explanatorily interesting. The second, which lies at the heart of the present discussion, is to which degree should we preserve current taxonomies of psychological phenomena. On one hand, we may think that we must preserve emotion taxonomies as they are and attempt to map them onto the brain in whichever form. On the other hand, we may think that if mappings between emotion categories and brain regions or networks fail to obtain, we should revise emotion categories themselves. Answering this question requires an account of what are the explanatorily relevant divisions in which we can divide the brain, be it at the level of anatomical regions, intrinsic networks, or functional networks.

Regarding these questions, Anderson (2015) distinguishes three approaches:

Conservatives It should be possible to specify a set of fundamental operations that will allow cognitive theories and process models to map more cleanly onto the brain than is currently evident. We should consider neurobiological evidence during the analysis and decomposition of a cognitive process into its components, rather than perform this in isolation.

Moderates Many elements in the current ontology may not be composites, and some elements may not reflect any aspect of psychological reality. A careful examination of the way the brain responds during experiments designed to manipulate these constructs can reveal them for what they are, suggesting splitting and lumping categories when the brain calls for it. At the end, we will have a set of mental operations that will often map to brain regions one-to-one.

Radicals We should rethink the very foundations of psychology. The required revisions may involve the construction of categories that divide the mind in very different ways than the current ontology does. At the end we may have very few one-to-one mappings between brain regions and psychological primitives. (adapted from Anderson, 2015, p. 70)

In the case of emotions, conservatives would argue that there must be a one-to-one mapping between emotion categories and brain regions. Moderates, in turn, would be open to revising emotion categories depending on empirical findings about the brain. Lastly, radicals argue that we should abandon emotion categories if necessary if findings about how the brain is divided call for it.

I will not attempt to solve general questions about cognitive ontology here. Yet, I will venture the following position. In my view, there are good reasons to maintain emotion categories, since they constitute the explananda of emotion research. Many of the examples championed by defenders of radicalism are constructs which are posited in the context of scientific psychology, such as working memory, or attributions such as attributing the left lateral fusiform area the function of visual word form processing (Price & Friston, 2005). These latter constructs are posits meant to explain other phenomena of which we do have a clear idea, namely, short-term memory or reading. The phenomena themselves are not up for elimination, since they constitute what psychology is ultimately trying to explain. I believe that emotion categories resemble short-term memory and reading more than they resemble working memory or visual word form processing as a proper function of the left lateral fusiform area.

On the surface, this would appear as a defense of a conservative position. There is a sense in which I do endorse such a view. Specifically, I believe that we cannot dispense with emotion categories, since doing so will lead to a change of subject. In other words, we cannot explain emotions without some recourse to emotion categories (which I will explain in detail in chapter 5). Yet, if we approach the issue from the perspective of brain function, I endorse a more radical view. As it will be clear in the second part of this work, particularly in chapter 6, we should accept that emotions can be multiply realized in a number of brain regions and networks. If this is the case, then it may be true that the brain does not mirror distinctions in terms of emotion categories. In other words, whatever the brain does, whatever its proper functions may be, they may not map onto emotion categories. However, I do not take this to be sufficient to eliminate emotion categories, as I will argue in later chapters.

In any case, resolving the issue of how to individuate neural patterns requires raising questions, not only about correspondences with emotion categories, but also about

what a scientifically tractable taxonomy of brain function will look like. As it stands, it seems that the best prospects for defenders of coordination at the neural level is to appeal to pattern assignment. This is not to say that the pattern assignment will be ultimately successful, as future research may show its limitations. All that I claim here is that if scientists are to look into the brain for correspondence and coordination, the current best candidate strategy is to assign emotions to overall patterns of activation rather than locating them in specific and consistent brain regions or intrinsic networks. In turn, this requires a defense of functional networks as explanatorily relevant. Without such an account, we have good reason to accept that coordination at the neural level fails to obtain, and thus that $\text{NOC}_{\text{Neural}}$ is a well supported claim.

2.2.2 Physiological patterns

Physiological evidence concerns the presence or absence of patterns of autonomic nervous system activity (hereafter ANS activity). ANS activity measures can be divided into three categories. First, there is activity related to the cardiac system, which includes heart rate variability (HRV), blood pressure, cardiac cycles, and the like. Second, we find variables regarding respiration, e.g. respiratory cycles, respiration period, amplitude, etc. Lastly, there are variables concerning electrodermal activity, i.e., skin conductance levels, responses, resistance, etc.

Given these types of physiological variables, whether or not there are coordinated patterns of physiological activity can be broken down into two criteria. One is determining whether there is patterning within a class of variables. We can ask whether there are specific patterns concerning cardiac, respiratory, or electrodermal activity for a particular emotion. Additionally, we can investigate patterning between the classes of variables, i.e., whether cardiac, respiratory, and electrodermal activity are robustly correlated and form a homogeneous set of responses for each emotion.

In an early study on autonomic activity associated with emotions, Ekman, Levenson, and Friesen (1983) investigated surprise, disgust, sadness, anger, fear, and happiness, the emotions they included in their earlier basic emotions lists. In one task, they asked subjects to contract specific muscles in order to implicitly mirror a given facial expression, without telling them explicitly which expression it was. The idea behind this task is that, in their view, facial expressions corresponding to a given emotion would elicit the corresponding autonomic markers. In another task, the researchers asked subjects to relive a past experience that would elicit a given emotion (this time explicitly telling them which emotion it should be). During these tasks, the investigators measured the subjects' heart rate, left- and right-hand temperatures, skin resistance, and forearm muscle tension (to control for heart rate effects due to subjects clenching).

Ekman et al. report a main effect of emotion in autonomic variables, i.e., that autonomic variables change significantly depending on the emotion. However, they also noticed an interaction between task and emotion. This means that there was an effect on autonomic variables depending on the task and the emotion elicited. Regarding the emotions themselves, they found that heart rate and temperature increased for

anger, as well as heart rate increases for fear, in contrast to happiness. They also hold that they were able to distinguish disgust and anger from each other and from fear or sadness in the first task, and that sadness from disgust, anger, or fear in the second. Lastly, the researchers claim that they could differentiate between negative and positive emotions in both tasks.

In a later report (Levenson, Ekman, & Friesen, 1990), the same researchers repeated this experiment in four different samples. Again, they report autonomic specificity for all emotions, particularly in terms of heart rate and skin conductance.[26] Regarding heart rate, anger, fear, and sadness elicited the greatest acceleration, followed by disgust, happiness, and lastly surprise. As for skin conductance, fear and disgust produced the most conductance when compared to the other emotions (which did not differ significantly from one another).

These studies, on the surface, suggest that there are autonomic patterns corresponding to at least some emotions. Yet, they fail to provide conclusive evidence. First, asking subjects to mimic a facial expression without explicitly naming the corresponding emotion does not prevent subjects from easily realizing the expression they are mimicking. If this is true, it is plausible that subjects finding out about a given emotion would trigger associations that would elicit the emotion via memory and imagery. Even though Ekman and colleagues report differences between the imitation and the memory tasks, these may have been due to subjects feeling nervous in either scenario or being distracted by their efforts to move individual muscles (which they do not move that way in a natural context).

But even if we accept that the tasks worked as they should have, still the distinctions and correlations they report do not provide a conclusive case for autonomic specificity. On one hand, the first report present an interaction between task and emotion, an interaction that is not explored further. How can we tell, then, that the effects they observe are evidence for physiological specificity instead of artifacts of the tasks at hand? Moreover, many of the expected differences fail to be significant. For example, fear and anger both share accelerated heart rate and skin conductance responses. At best, the researchers were only able to distinguish one positive from four negative emotions.

Later studies have attempted to expand on these findings and find better distinctions in terms of autonomic activity. Collet, Vernet-Maury, Delhomme, and Dittmar (1997) for example used slides portraying expressions for each of the emotions tested by Ekman and others, and asked subjects to feel the emotion corresponding to each of the slides shown by thinking about past experiences. This method of autobiographical recall mirrors again the one used by the other researchers. This time, however, Collet and colleagues measured a different set of autonomic variables, namely, skin resistance, skin conductance, skin potential, skin blood flow, skin temperature, and instantaneous respiratory frequency.

[26] Even though the previous study investigated skin resistance, the two measures are essentially the same, as they are complimentary.

Collet and colleagues report finding specific patterns for each of the emotions they studied. For instance, happiness is characterized by higher skin conductance and relatively low skin temperature, whereas anger displays high conductance and temperature. As they report their findings, none of these patterns is identical to one another, thus providing evidence for autonomic specificity. Additionally, they stress that their results go beyond the previous studies in that it provides distinctions between particular emotions and not only between positive and negative ones.

More recent studies have introduced multivariate techniques to look for physiological patterns. Rainville, Bechara, Naqvi, and Damasio (2006), for example, used principal component analysis to see which variables were the most useful when classifying different emotions from data on autonomic responses. As in other studies, they used autobiographical recall methods to elicit anger, fear, happiness, and sadness. They measured a number of variables regarding respiration and cardiac cycles, including respiration period, amplitude, heart-rate variability, and others.

In their analyses, the researchers first contrasted each emotion to a neutral condition independently. They report that the respiratory period decreased in fear and happiness and less consistently in anger, while the variability in respiratory period increased in sadness. Regarding heart-rate variability, there was a decrease within respiratory cycle and high frequency range were robust in fear and significant but weaker in happiness. Univariate comparisons between the four emotions showed that several dependent variables were sensitive to emoton-related effects. Indices of HRV and respiratory activity showed highly significant effects of emotions. Graphical representations contrasting pairs of emotions showed that anger was clearly separated from both fear and happiness based in respiratory period and respiratory rate.

Using exploratory PCA showed five contributing factors that explained 91% of overall variance. Factor 1 explained most of the variance in HRV measured within respiratory cycles. No respiratory variable loaded noticeably on Factor 1, which the researchers take as implying that this factor mostly captures HRV. Factor 2 corresponded mainly to respiratory period and part of respiratory amplitude. However, it also captures some HRV variance. Hence, the researchers conclude that factor 2 captures most of the variance in HRV coupled with respiration. In turn, factor 3 captured variance in mean respiratory rate levels independently of HRV and other respiratory variables. Factor 4 was associated with the frequency index of respiration variability, and lastly, factor 5 reflected variance in respiration period and amplitude. What these PCA analyses mean is that variance in physiological measures can be reduced to these five main factors. In other words, it is these groups of variables that distinguish among different emotions in terms of autonomic activity measures.

The researchers claim that this study provides some evidence that basic emotions are associated with distinctive patterns of cardiorespiratory activity. Different emotions were distinguished from a neutral condition based on different subsets of dependent variables and multi-dimensional exploration of the data revealed complex patterns of activity that characterized each emotion. According to the PCA, the variance in cardiorespiratory activity can be explained along five dimensions, mostly HRV.

Besides these particular studies, seen from a panoramic perspective, some meta-analytic findings also support the case for autonomic specificity. Kreibig (2010), for instance, covered 134 publications and examined three classes of variables: cardiovascular, respiratory, and electrodermal. She reports specific patterns for a great number of emotions. For example, she claims that anger involves faster breathing as seen in shortened inspiration and expiration times, more expiration than inspiration, increased heart rate, increased overall blood pressure, among others. Fear elicited a similar pattern, involving broad sympathetic activation, cardiac acceleration, increased vasoconstriction, and increased electrodermal activity. However, in the case of fear, peripheral resistance decreased whereas it increased for anger.

Other reported findings show that we might even get more fine-grained categories by looking at physiology. The case of sadness is a prime example. Kreibig reported two types of physiological responses for sadness: activating and deactivating. Activating sadness, or crying sadness, is characterized by increased cardiovascular sympathetic control and changed respiratory activity. Specifically, it is correlated with increased heart rate and increased skin conductance levels. Deactivating sadness, or non-crying sadness, is characterized by sympathetic withdrawal, and decrease in electrodermal activity, as seen in a decreased heart rate, longer preejection period, increased heart rate variability, decreased diastolic blood pressure.

Despite the studies supporting autonomic specificity, there are also efforts trying to challenge it. In a recent meta-analysis, Siegel et al. (2018) evaluated empirical evidence in favor or against specificity. In their own words, they contrasted the hypotheses that there is "limited ANS variation around a fingerprint (the classical view)" against the hypothesis that there is "substantial variation that is meaningfully tied to the situation (the constructionist view)" (Siegel et al., 2018, p. 347). In contrast to previous studies and meta-analyses, they claim that evidence supports the latter hypothesis.

The authors included 204 studies from 1950 to 2013, and studied the same three-fold division of variables used by Kreibig, namely, cardiovascular, respiratory, and electrodermal measures. To do this, they compared the effect sizes across all their studies. Some results display mean ANS changes from baseline across several effect sizes but with substantial variability. For instance, the patterns of anger and fear showed large effect sizes, suggesting that their physiological patterns differed significantly from baseline across several autonomic variables (specifically, heart rate, cardiac output, diastolic and systolic blood pressure). Yet, these effect sizes are very heterogeneous, indicating that even though these emotions have clear physiological effects, these effects are not uniform and do not form a stable pattern.

Other results show small mean ANS changes and moderate variability. For example, the researchers report the cases of disgust and neutral categories. For disgust, only skin conductance level and responses had relevant effect sizes, but only the latter was homogeneous. For neutral conditions, only systolic blood pressure had an interesting mean effect size, but it is also a heterogeneous variable. Happiness and sadness had increased effect sizes in heart rate, diastolic blood pressure, skin conductance level, and others, but they are mostly heterogeneous. As a result, most mean ANS changes

were not uniform. Additionally, the researchers claim that ANS changes were not specific to a given emotion category either. Happiness had a mean increase in skin conductance level similar to disgust, anger, fear, and sadness, for example.

We can now interpret evidence challenging physiological specificity using the categories presented above. On one hand, there is evidence suggesting that there is low coordination between cardiac, respiratory, and electrodermal variables. Evidence of the first type presented by Siegel et al. is one example. As they suggest, correlations between physiological variables preclude their classification as a specific pattern. On the other hand, there seems to be evidence showing low coordination within physiological variables, as presented in the second group of findings reported by Siegel et al. According to this argument, some physiological variables have more impact than others in determining the ensuing emotion. As a result, given the lack of correlation between physiological variables, the researchers claim that there is no physiological specificity for emotion.

Nevertheless, settling this discussion requires further methodological and epistemological decisions. First, it is unclear whether all physiological variables should have the same influence when considering whether there is a coordinated pattern or not. Often used variables such as heart-rate variability surely are among the most important ones to consider. Yet, the status of other variables such as respiration period or vaso-constriction is left undecided.

To determine the relevance of such variables for the purposes of emotion classification, we must be clear, first, about their causal relationships, and second, in case where no causal relationships hold, how we expect them to correlate. For example, it is plausible that a number of cardiac variables are causally connected with electrodermal variables. An account of physiological patterns should take these connections into account when classifying autonomic activity into sets of responses. If these causal connections fail to hold, but there is correlation, we need to be clear what these correlations entail and whether they are robust enough to warrant classification under a pattern.

Second, both optimistic and skeptical researchers fail to distinguish between within- and between-variable coordination. This leads to an ambiguity that affects both camps. On one hand, it could be the case that we need correlations among variables of the same type (say, respiratory variables) in order to consider that there is a robust physiological pattern (i.e., a respiratory pattern). On the other hand, we may not demand correlations within a given family of variables, but rather between some measures of different types. For instance, we may expect some cardiovascular measures to correlate with some electrodermal ones, without the requirement that there are within-variable patterns. As it stands now, researchers highlight evidence showing that one measure is associated with a given emotion or that another is not, without a clear argument as to whether it is necessary that all measures of a given type correlate with one another or whether it is necessary that some measures of different types do so.

Lastly, similar to the discussion regarding neural patterns, the use of multivariate techniques is still controversial. Presumably, lots of physiological processes obtain when we experience an emotion or any other state. As a result, whether or not the ability to classify them with analyses such as PCA tell us something explanatorily relevant remains unclear. Skeptics may argue that the mere presence of a statistical pattern says little about the causal mechanisms involved in emotion. Optimists may react by pointing out that multivariate techniques are nevertheless more robust and that they do not claim that there is just any pattern at play. In any case, we need criteria to count statistical relations between variables as dividing physiological responses into distinct types. As long as researchers do not agree on these criteria, discussions about physiological specificity are bound to remain indecisive.

2.2.3 Behavioral patterns

Behavioral patterns refer to possible behavioral outcomes of an emotion episode. In the current literature, the best account of the behavioral patterns of emotion comes from appraisal theories. According to these theories, emotions involve states of action readiness (Frijda, Kuipers, & ter Schure, 1989). On one influential construal, action readiness is cashed out as the individual's readiness or unreadiness to engage in interaction with the environment. This may consist in readiness to engage or disengage from interaction with some object in a particular way (action tendency) or in a general state of activation or inhibition of behavior (activation modes) (Frijda, 2007). Among the major modes of action readiness, as Frijda calls them, we find "moving toward," "moving away," "moving against," "[being] helpless," "submission," "rest," "[being] in command," "[being] excited," "apathy, disinterest," and "undo" (Frijda, 2007, p. 34).

Some researchers claim that we can differentiate between emotions by appealing to the different states of action readiness they elicit. On one such study, Frijda et al. (1989) asked subjects to recall instances of emotions and asked them to rate different statements concerning various action patterns. These statements included descriptions such as "I wanted to approach or make contact" or "I wanted to oppose, to assault." They then tried to map patterns of action to emotion names by investigating how well they could predict the emotion label from these patterns. Frijda and colleagues report some predictability for 32 emotion categories. Among the highly correlated patterns they report crying for sadness, protecting one self for fear and anxiety, moving against an object for anger, avoidance for disgust, and hiding from others for shame, among others. This suggests that there may be some correspondence between action readiness states and emotion categories.

Other studies have yielded similar results. Roseman, Wiest, and Swartz (1994) used as similar experimental design, asking subjects to recall past emotional experiences, narrate them, and answer a questionnaire that tapped into their behavioral outcomes, among other variables. Questions relating to behavioral outcomes included questions about whether during a given emotion, they wanted to approach something or avoid it, or whether they wanted to cry or resign to something, and the like. The researchers claim that their experiment shows clear distinctions between 10 emotions in terms of

their action tendencies. Among the tendencies reported, we find fear as a readiness to reduce the possibility of harm, sadness as crying and seeking comfort, disgust as attempting to get something noxious out of the body, among others. The emotions they claim they could distinguish were fear, sadness, distress, frustration, disgust, dislike, anger, regret, guilt, and shame.

In spite of these optimistic efforts, cashing out emotions in terms of action readiness and action tendencies does not go without problems. On one hand, there is some observed variability. In Frijda et al. (1989) we find one action pattern corresponding to two emotions (e.g. protecting oneself in fear and anxiety). Frijda (2007) recognizes this, and writes:

> Major mode [sic] tend to map on major emotion categories. "Fear" corresponds with a tendency to move away, "anger" with the tendency to move against (or "oppose") as well as "hurt"), "sadness" to being helpless, and so forth. The modes *are not truly linked to particular emotion categories*, though. Different emotions may share the same action tendency, as do timidity and fear, and humility and shame. Different instances of one emotion class may differ in action tendency; so, for instance, "anger out" and "anger in." (Frijda, 2007, p. 34; emphasis added)

As it should be clear from the preceding quote, Frijda's view regarding the role of action tendencies as distinguishing between emotion categories is not as optimistic as it may have appeared at a first glance. He thinks that some action tendencies to map onto emotion categories, but he does not think that action tendencies are sufficient to individuate emotions. As a result, Frijda seems to endorse the claim that there is no correspondence between emotion categories and behavioral patterns, i.e., $\text{NOC}_{\text{Behavioral}}$.

On the surface, this would only mean that variability in terms of behavior is still controversial. However, the problem runs even deeper. Critics of appraisal theories have argued that the links between emotions and action tendencies may as well be a matter of conceptual truth rather than empirical fact. If so, questions about correspondence become trivial; emotions will trivially correspond to behavior patterns (just as water "corresponds" to H_2O.) Consider the presumed correspondence between fear and engaging in behavior towards protecting oneself in situations of perceived harm. Suppose we attempt to falsify such correspondence. We would need to be able to obtain a fear state that does not involve such a behavioral tendency. Yet, arguably, that tendency is precisely what it means to be in a fearful state. As a result, any candidate state to falsify this supposed hypothesis would not count as a fear state as a matter of conceptual fact.

This problem can be brought to light by considering moves to ameliorate it. Roseman (2011), in response to challenges of variability between emotions and behavioral outcomes, argues that emotions are consistent at the level of coping strategies. He claims, for instance, that fear forms a consistent pattern insofar as it involves a strategy to move away from or stop moving toward some danger. Again, we could ask:

what would it mean for this to be false? Presumably, this correlation obtains, not as a contingent fact, but because the behavioral outcome provides a definition of what it means to be afraid. To use Smedlund's (1992) example, these results are as if we discovered that bachelors are male and single.[27]

A similar worry runs regarding LC in the case of inter-domain coordination. Depending on how we carve out behavioral outcomes, any outcome that may correspond to a give emotion, even if not one-to-one, can be spelled out to yield a correlation with some neural and physiological state. If we carve out behavioral outcomes in a fine-grained fashion, coordination between the neural and physiological domain would be almost trivially true. One can resist this result by clarifying that triviality only obtains if neural and physiological states are interpreted as token states, not as types, i.e., by adopting a coarser grain. Still, the question of how to spell out these patterns properly remains unanswered.

In chapters 5 and 6, I will suggest a way out of this difficulty in detail. In my view, we can rely on folk psychological vocabulary to determine the grain to carve out behavioral outcomes, by noticing distinctions in how we distinguish emotions in our everyday concepts. We can then explicate folk emotion concepts into more abstract, functionally defined scientific concepts of emotions that yield empirical hypotheses regarding dispositions to engage in certain behaviors which individuate emotions. For the time being, let it be clear that as the literature stands right now, there is no clear account of how to individuate behavioral patterns in an empirically interesting manner. This makes claims about coordination and correspondence in the behavioral domain problematic. As I will argue at the end of this chapter, behavioral evidence is important for emotion research, but requires further conceptual work. Before I dwell into these arguments, let us go on to the next type of pattern involved in VT, namely, expressive patterns.

2.2.4 Expressive patterns

One of the main issues surrounding the existence of emotions as distinct, universal constructs come from the main source of evidence motivating traditional basic emotion theories: the universality of facial expressions. As I explained in the previous chapter, universality was an important piece of Ekman's and Izard's versions of BET and still remains a contentious issue.

Regarding the issue of emotions as kinds, the question of universality can be framed as a question about the historical, functional, or social nature of emotions. In traditional versions of BET, universality entails that emotions are evolved and linked to biologically patterns of autonomic and neural activity. In turn, opponents of BET claim that if universality is not established, this would show evidence for the functional or social nature of emotions (Crivelli & Fridlund, 2018) or it would provide support for skeptical theses (Nelson & Russell, 2013; Russell, 2003).

[27] McEachrane (2009) makes a similar point. In his view, appraisal theory is pseudoempirical since it individuation in terms of appraisal or action tendencies follows from the meanings of the emotion terms involved.

In order to explore the issue of universality, it is paramount to define first what universality means. Russell (1994) distinguishes four propositions related to universality:

(a) Specific patterns of facial muscle movement occur in all human beings.

(b) Certain facial patterns are manifestations of the same emotions in all human beings.

(c) Observers everywhere attribute the same emotional meaning to those facial patterns.

(d) Observers are correct in the emotions they (consensually) attribute to those facial patterns. (cf. Russell, 1994, p. 106)

The first of these propositions relates to the specificity of the facial expressions themselves. The second, to their correspondence to emotional states. It is important to note the independence between these two. It is logically possible that there is a limited repertoire of specific facial expressions that does not match to the number of emotions that they can express. For example, suppose that the Duchenne smile is one of such specific expressions that occurs in all human beings. Yet, it may be possible (and in fact is often the case) that the Duchenne smile can express both happiness and pride. Conversely, there can be a fixed number of emotions but no specific patterns of facial movement that occur in all human beings. Facial expressions could vary from culture to culture such that there is no fixed repertoire of expressions, without the same applying to emotions themselves.

The third and fourth propositions relate to the observers of these facial expressions. Proposition (c) is descriptive, that is, it merely describes a possible state of affairs, a sociological and psychological fact about how we attribute emotions. Proposition (d), however, is normative, as it involves standards of correctness about the attribution of emotions via facial expressions. Furthermore, (d) requires establishing at least (c), since in order for observers to attribute emotions correctly everywhere, they must first attribute emotions in a uniform manner.[28] Since I am interested in questions about correspondence between expressions and emotions, and not about the correctness of attributions in terms of expressions, I will not deal with (c) and (d) at the moment. Instead, I will focus my efforts on (a) and (b), i.e., whether empirical evidence helps establish the existence of specific facial movement patterns and their correspondence to emotions.

The main defender of universality is, without a doubt, Ekman (see e.g. Ekman, 1972, 1980; Ekman & Friesen, 1971; Ekman et al., 1987, 1983; Ekman, Sorenson, &

[28] Russell thinks that it requires all of the previous three propositions. However, it is possible, albeit unlikely, that observers everywhere attribute the same emotional meaning to expressions even in the absence of a specific repertoire of facial movements and a fixed number of emotions. This would require observers to be right every time the see an expression, even if they have not seen it before, and to attribute the right emotion, even if they have not encountered it in the past. As unlikely as this may be, it is logically possible.

Friesen, 1969). I will not comment all of these studies in detail, but I will expand on two groups: the series of American-Japanese studies presented in Ekman (1972) and the study in New Guinea in Ekman and Friesen (1971). Not only are these studies some of Ekman's most cited, but also great examples of Ekman's groundwork for his theory.

Ekman (1972) presents a series of studies (prior to Ekman and Friesen (1971), which I will discuss below) in which he investigated how similar were the facial expressions and their recognition among Japanese and US Americans. In the first study, he compared how well can one culture recognize emotions in the expressions of the other culture as well as their own. According to Ekman, if emotional expression were culture-specific, recognition would be high for one's own culture but not for others.

In this initial study, Ekman showed both Japanese and US American samples stress-inducing and neutral videos. They then videotaped these reactions and showed them to four separate groups in Japan and the United States. Viewers of these reactions were instructed to judge whether the person in the video had watched the stressful or the neutral film. As he expected, Ekman found that both Japanese and US American samples were capable of recognizing expressions in both cultures above chance (around 60% of the time). This was, in his view, evidence that emotional expression was universal.

In spite of his optimism, the study suffered from three important flaws. First, the experiment only showed that whatever expression both Japanese and US American samples showed, they are similarly interpreted by both cultures, but not that they had actually portrayed the same expressions. Second, the study can not say anything about whether expressions are specific to each emotion or not, since the only conditions tested were stress (unpleasant) vs. neutral. Third, it could not rule out learning effects due to exposition to visual representations of the other culture (e.g. in TV, magazines or books).

To address the first two problems, Ekman repeated the previous study—again between Japanese and US American samples. This time he measured the participants' facial expressions using an instrument designed along with Friesen and Tomkins (Ekman, Friesen, & Tomkins, 1971), the Facial Action Scoring Technique, FAST (which would later become the Facial Action Coding System, FACS). FAST consisted in dividing facial expressions into different muscle movements, determining the beginning and end of each movement, and classifying them according to a predefined list of items such as raising the eyebrows or opening the mouth. By using FAST, Ekman and colleagues could compare expressions between the different cultures and determine whether the expressions were similar themselves and not only judged in the manner. Additionally, they tested for six of Tomkins's primary affects: Surprise, fear, anger, disgust, sadness, and happiness.

In this new study, Ekman reported obtaining correlations between each culture's facial expressions at the level of each item (e.g. they raised the eyebrows at the same time). Additionally, when the researchers categorized expression items into emotions, correlations between these emotions were higher. That means, when they classified an

expression as a sadness expression, they found that instances of sadness were correlated in the two cultures. The same logic then applied to the other primary affects they tested. This would presumably show that not only were expressions in the two cultures similar, but also that they corresponded at the level of specific emotions.

These studies provided a basis to think of emotions or at least emotional expression as universal. In later studies (see Ekman, 1972), Ekman would reportedly replicate and extend these findings to other cultures and with other methods. There was, however, a persistent problem that would be addressed in a study years later, namely, that there was no way to control for exposition and learning effects. Most of the cultures Ekman tested had been in one way or another subject to the representations of emotions that might have biased the results. As a result, Ekman opted to conduct a study in a remote culture that would have had no contact with Western cultures and therefore would prove a true test of universality.

Ekman and Friesen (1971) conducted a study with members of the Fore group in New Guinea. The Fore were isolated until the mid-twentieth century, hence they constituted an interesting community in which to test Ekman's hypotheses. The researchers showed subjects three photographs of different facial expressions and told them a story. Subjects then chose the photograph that matched the story's emotional content. Stories included content for happiness, sadness, anger, surprise, disgust, and fear. These six emotions would make up one of the most widely used lists of basic emotions.

In their report, Ekman and Friesen showed that subjects were generally able to identify the correct photograph at a high success rate. In their view, this result provided evidence that there were facial expressions that were universally recognizable, even in cultures with no contact with Western societies. In later years, Ekman and colleagues (1987) would use this method in other Western cultures and replicate these findings.

Against universality

In spite of Ekman's optimism, universality does not go without its critics. Arguments against universality come in three main strands. The first strand of criticism intends to cast doubt on the robustness of the findings presumably supporting universality, showing flaws in the designs as well as the assumptions of a number of studies. The second strand tries to outweigh empirical evidence for universality by underscoring cultural variation.

Methodological criticism

The most influential methodological criticism of universality comes from Russell (1994).[29] In his review, Russell argues that universality studies are plagued with problems in their ecological, convergent, and internal validity. If Russell is right, it is doubtful

[29] Some of the arguments in this review are rehearsed and updated in Nelson and Russell (2013).

how much we can trust previous findings, let alone establish correspondences between expressions and emotions. Let us go through some of his arguments at a time.

First, Russell argues that there is little information on what facial expressions occur naturally in the societies studied. On one hand, a number of studies have used highly artificial stimuli whose ecological validity is unclear. These stimuli are mainly posed facial expressions, which are seldom present as such in natural contexts and, moreover, already presuppose which expressions are to be found. This is because actors posing these expressions are already taught how these expressions should look like. Also, study designs using these sets already presuppose that emotions are divided into the number of expressions available in the set.

On the other hand, designs involving forced-choice paradigms may guide subjects into specific responses. This is problematic, for example, in both of Ekman's studies mentioned above. If researchers ask subjects to pick from a fixed set of labels which emotion does an expression signal, subjects are bound to pick the label that mostly applies in that case, even if they would categorize it differently in other contexts. Since we lose contextual information when forcing our choices to a fixed set of labels, studies using these designs incur in issues with their ecological validity.

Additionally, Russell reports that when participants are allowed to freely choose their labels, agreement between them drops dramatically. To make matters worse, even if we used free labels, the interpretation of the results is made difficult by problems regarding synonymy and translation. Suppose we run a study in two different cultures with two different languages. We obtain then two lists of possible emotion terms, and we want to know whether these two lists match up. To do so, we must translate the terms of one list to the other. Even if they don't correlate one-to-one, it is difficult to say whether a given translation is adequate without presupposing already our own categories. This thus becomes a radical translation problem, one that may plague studies overall.

Second, Russell attacks the internal validity of these studies, claiming that there are a number of factors that researchers have not controlled properly that may introduce a number of confounds. For example, in some studies (e.g. Winkelmayer, Exline, Gottheil, & Paredes, 1978), participants are shown the whole set of pictures before rating. This already gives subjects information about which expressions they are going to see, and may already lead to pre-categorizing according to the number of different expressions in the set.

In addition, problems regarding possible learning effects or familiarity with experimental hypotheses are frequent. One such example regards the use of college students as participants. College students may be exposed already to the categories and expressions that these studies undertake to obtain. This may happen either directly, for instance, by using psychology students, or indirectly, by students being exposed to representations of emotions in terms of fixed facial expressions in popular and artistic media.

Lastly, problems with convergent validity, Russell claims, include the use of similar methods in most studies and a lack of methodological variation. In his view, studies

mostly used what he calls the *standard method*: subjects are shown preselected still photographs of largely posed facial expressions and then asked to choose one of a fixed number of alternatives. At best, Russell explains, studies use variations in one or two elements of the standard method (e.g. varying the length the lists, using free labeling, using films rather than photographs). Yet, he argues, little is known outside of these small variations.

Empirical criticism

On the side of empirical evidence, attacks on universality come from two sources. One is evidence showing that agreement among cultures regarding which facial movements correspond to which emotions has been overstated. Rather than finding robust agreement, researchers have showed that agreement drops under certain conditions. The other source stresses cultural variation, showing differences between different populations in terms of their perception and categorization of facial expressions.

One example of the first source of evidence is the meta-analysis by Elfenbein and Ambady (2002). Using the same data as that in Ekman's and Izard's studies among others, Elfenbein and Ambady show that there is an in-group advantage in facial expression recognition. In other words, the researchers show that members of the same group are more accurate in judging expressions of members of their same group. Specifically, they report that Western participants are 9.3% more accurate when judging other Western faces than with African or Asian ones. Even if overall recognition is still above chance level (58% accuracy), these results suggest that there is important accuracy scores depend on culture and are not as uniform as defenders of universality might think.

Examples of the second source of empirical evidence involve studies in a number of cultures. Elfenbein and Ambady themselves claim that when analyzing data in terms of individual emotions, some emotions are poorly recognized universally. In their meta-analysis, they found that fear and disgust are the most poorly recognized, even though they are among the most cited candidates to universal, biologically basic emotions. According to them, this implies that culture still shapes meaning of faces even in the presence of some uniformity.

Other studies also show mismatch between Western and non-Western interpretations of faces. Crivelli and Fridlund (2018) report that communities from the Trobiand islands in New Guinea understand gasping faces as threat displays instead of fear displays, as traditional studies have attempted to show. Similarly, Gendron, Roberson, van der Vyver, and Barrett (2014) found that the Himba people of Namibia perceive facial actions in context, that is, not as corresponding to a feeling but to the whole situation. For example, instead of interpreting crying as corresponding to a feeling of sadness, they situate it as a response to death, showing differences in the intentionality of their interpretation. Jack, Garrod, Yu, Caldara, and Schyns (2012) report that East Asian facial expressions overlap considerably, leading to fuzzy categorization contrasting with Western taxonomies. Along the same lines, Jack, Sun, Delis, Garrod, and

Schyns (2016) claim that in Chinese societies categorize emotions into more categories than English samples when asked to judge facial expressions.

Variability and expression

Universality, without a doubt, remains a controversial subject. In spite of Ekman and others' optimism, evidence for universality is still inconclusive and has been challenged in a number of ways. However, given that the heaviest criticism is methodological, we cannot yet rule out some degree of universality. There is some evidence supporting universality, albeit a weak interpretation.

Let us go back to Russell's propositions concerning universality. Regarding the first proposition, that specific patterns of facial muscle movement occur in all human beings, evidence seems to speak in its favor. Not only is there some recognition of the same set of pictures across cultures, but also more recent studies have shown some patterns that obtain reliably. For instance, Jack et al. (2016) report that four patterns described in terms of action units (individual muscle movements in the face) obtain and are recognized in both Western and non-Western samples. In the same vein, Cordaro et al. (2018) report patterns for 22 emotions in five cultures (China, India, Korea, Japan, and the US). In this sense, we can take the first proposition to be partially supported by empirical evidence.

Yet, when it comes to the second proposition, that certain facial patterns manifest the same emotions in all human beings, matters are still unclear. First, the methodological flaws that Russell has pointed out suggest that a number of studies suffer from problems with circularity, since the categories and correspondence they set out to prove already inform subjects' responses, and hence are already implicit in the experimental designs. More importantly though, issues with translation and the intentionality of facial expressions cast doubt on the correspondence between these and emotional states. Without a common background on which to judge the content of facial expressions, there will be a constant mismatch between how different cultures read out the face.

Evidence for cultural variation in emotion expression production and recognition tries to dismantle the second proposition presented above, namely, that emotional expressions are manifestations of the same emotions in all humans. On the face of it, there could be universal patterns of expression (assuming the methodological criticism is misguided), but they do not correspond to the same emotions everywhere. If this is true, then we would have evidence to reject LC, but also we would have to accept $\text{NOC}_{\text{Expressive}}$.

Apart from showing how universality is controversial, I suspect these findings suggest that the question of the universality of emotional expression is a different topic altogether that does not have much bearing on the issue of variability. Even if there were no universal patterns of emotional expression (LC) and hence no correspondence with emotions ($\text{NOC}_{\text{Expressive}}$), would that entail that emotions are variable phenomena? Plausibly not. The reason is that there could still be fixed, even innate patterns at the neural or physiological level that would grant emotions some robust form of ho-

mogeneity. Otherwise, we would be forced to split emotion categories in terms of their different expressions even in presence of evidence for neural and physiological homogeneity. If this line of argument is correct, it follows that evidence on the universality of emotional expression is largely irrelevant to the question of variability.

Someone may object that expressions are still a central part of our emotion attribution and behavioral manifestation. If this is so, the objection would hold, we cannot set evidence on expression aside, as it provides important information about how we individuate emotions folk-psychologically. In my view, this objection leads to an interesting consequence: it reduces expressions to subsets of behavioral patterns. According to the objection, expression matters because they form part of behavioral outcomes that help distinguishing between emotions. If this is true, there would be no reason to consider expressive patterns as separate from behavioral patterns (action tendencies) altogether. In other words, expressions are at best part of our action tendencies, and at worst irrelevant to the case of variability.

2.2.5 Phenomenological patterns

Lastly, we come to phenomenological patterns. Phenomenological patterns are often understood as patterns of subjective experience. What exactly characterizes subjective experience is nevertheless unclear. Subjective experience, taken as a criterion to individuate patterns candidate to correspondence and coordination, fails on two grounds. First, the phenomena subjects and theorists describe as subjective can plausibly be reduced to other types of patterns already under consideration, such as collections of neural and physiological states as well as action tendencies (behavioral patterns). Second, even if there is some remainder in terms of qualitative experiences, there are good reasons to doubt these can successfully help us individuate emotions, thus precluding claims even about their variability.

A first approximation to tap into phenomenological patterns of emotions is to rely on self-report data, asking subjects to narrate or describe their own emotional experience. One influential example of such an approach is the work by Davitz (1969). Davitz undertook to develop a dictionary of emotions that synthesized how people use language to refer to their own emotional states. Based on a short list of emotion terms (Affection, Anger, Anxiety, Boredom, Cheerfulness, Confidence, Impatience, Sadness, and Satisfaction), he interviewed people asking them how they would describe each state and recorded their reports. To this list of statements he then added more descriptions from 1200 subjects who were asked to think of concrete instances of each emotion. From these reports Davitz obtained a list of 556 statements about emotion experience. Lastly, he asked a third group to rate how adequate each statement was to describe their own experiences. With this material in hand, he compiled the most used descriptions for each term into the dictionary.

A short examination of the definitions and statements found in Davitz's dictionary shows that many of the descriptions presumed to tap into subjective experience can be reduced to other patterns. For example, 'anger' includes among its most common descriptions 'my blood pressure goes up,' 'my pulse quickens,' or 'my heart pounds.' In

the case of sadness, we find 'there is a lump in my throat,' 'there is a clutching, sinking feeling in the middle of my chest,' and 'I have no appetite.' Similar descriptions can be found for other emotions terms as well. In these cases, it is easy to see that these patterns can be described as physiological patterns corresponding to each emotion. Other patterns in Davitz's dictionary's entries refer rather to action tendencies. In the cases above, anger includes 'my fists are clenched,' 'there is an impulse to hurt, to hit, or to kick someone else,' and sadness, 'tears well up' or 'I cry.'

Perhaps a more sophisticated approach to the phenomenology of emotion is found in Lambie and Marcel (2002). Lambie and Marcel distinguish three types of questions regarding the empirical investigation of emotions: (a) what is the content of emotion experience as it is experienced?, (b) to what nonconscious process or representation does emotion experience correspond?; and (c) what processes or differences in content lead to and contribute to emotion experience? In their view, only the first of these questions tackles phenomenology.

To answer the question of what is the content of emotion experience, Lambie and Marcel separate between emotion states and emotion experiences. Emotion states are the functional aspects of emotion apart from conscious experience, which include primary appraisals of events in terms of relevance to the organism, the activation of brain and bodily systems, and preparation for action. Emotion experiences are both the phenomenological aspects of emotional states (first-order experience) and the awareness of these experiences themselves (second-order experience).

For the purposes of evaluating NOC and LC, following Lambie and Marcel, we would have to decide at which level are emotions individuated. Suppose we decide that they must be individuated at the first-order experience level. Characterizing first-order experience is problematic for a number of reasons. As Lambie and Marcel recognize, our first mode of access to first-order experience is by introspection, which requires awareness of it, which in turns changes the first-order experience itself. The authors suggest that we can instead rely on memory and episodic reinstatement to tap into previous episodes of first-order experience, circumventing this problem. Yet, there is good evidence on memory manipulation showing that memory is also affected by our current epistemic states (Brown & Marsh, 2008; Edelson, Sharot, Dolan, & Dudai, 2011; Loftus, 2005; Mazzoni & Memon, 2003). If this is the case, relying on episodic reinstatement does not fix the problem.

A second worry regarding first-order experience is that, as Lambie and Marcel have characterized it, it includes aspects that are again reducible to other patterns. Given that first-order experience has underlying brain and bodily states, as well as involving action tendencies and appraisals, it is unclear why this level of description would yield a different type of pattern at all. As the characterization stands, Lambie and Marcel have suggested facets of neural, physiological, and behavioral patterns that are involved in emotion, but have not shown that there is something uniquely phenomenological worth separating.

One may object that these reductions leave the qualitative character of emotion experience untouched. In the same vein as proponents of the explanatory gap in

philosophy of mind (Chalmers, 1997; Levine, 1983) one could argue that physical or behavioral states do not exhaustively describe pure forms of emotion experience. An argument of this sort seems to be in the background of LeDoux's (LeDoux, 2012, 2013; LeDoux & Brown, 2017) claim that emotions and feelings should be used interchangeably. As a consequence, LeDoux recommends not making reference to emotions when we talk about circuits underlying survival behavioral dispositions, and instead looking for a theory of emotional consciousness as a theory of emotion.

This approach would entail individuating emotions by their qualitative character alone. As the record of discussions on the hard problem of consciousness attests, problems soon arise. First, to do this we need an account of how we could ground emotion concepts on first-person qualitative properties. Since we presumably do not have access to others' first-person experiences, we seem to fall prey of arguments such as Wittgenstein's private language argument (Wittgenstein, 1953/2009). According to a broad and naïve construal of this argument, it is nonsensical to think that the meaning of concepts such as 'red' or 'yellow' (and in this case 'sadness' and 'fear') can grounded in first-person experience, since we would have no public criteria for their correct application. This leaves us with concepts on which we cannot construct a scientific theory.

Second, we could resist the private language argument (and other similar ones) and insist that there is no reason why it would be impossible to ground phenomenal properties on publicly available criteria. There are good reasons to doubt that reductions of the phenomenal character of consciousness are impossible in principle (see e.g. Pauen, 2017), hence opening the door for third-person descriptions. In other words, we can reject the explanatory gap and defend the possibility of describing phenomenality in functional terms. This is for instance what LeDoux and Brown (2017) attempt in offering a theory of emotional consciousness. If this were so however, then we would be able to describe phenomenality in neural, physiological, or behavioral terms, thus diluting the category of phenomenological patterns into the other three.

Lastly, it is doubtful that the appeal to irreducible qualitative properties provides a tractable account at all. It is difficult to see, on their qualitative aspects alone, how emotions can differ from one another. This case is clear for emotions that are similar to one another like anger and indignation, or joy from pride (Prinz, 2007, p. 52). However, it is even more pressing for comparisons between intuitively very different emotions such as anger and fear, which are more similar to each other than, for instance, happiness and fear. Without invoking non-qualitative properties such as valence (which is ultimately a relational property), we cannot explain many differences between emotions (for an argument in this direction, see Frijda et al. 1989, p. 227).

Let us grant then that first-order experience cannot do the trick. We may still claim that phenomenological pattern individuation can obtain at the level of second-order awareness. Second-order awareness can be characterized, according to Lambie and Marcel, in two ways. In some cases, our emotion experience is directed to the self. These are cases where we, for example, experience anger as an offense to our own selves or sadness as an own failure. In other cases, emotion experience is directed

to the world. Here our emotions are describable in terms of objects, as when we experience the object of our anger as something offensive or blameworthy, or sadness as presenting a world that is unfulfilling.

The case for phenomenological pattern individuation at the second-order awareness level resembles attempts in appraisal theories to individuate emotions in terms of core-relational themes (Lazarus, 1991). It is also reminiscent of other approaches in phenomenology proposed by enactivists (Colombetti, 2009, 2017; Hutto, 2012). In both of these cases, second-order awareness refers to an experience of an emotion in terms of the relation between an organism and its environment, whether it is focused on the standing of the self as related to objects or focused towards properties of the objects as appraised by the self.

If this interpretation is correct, second-order awareness may be described in terms of other patterns as well, namely, as action tendencies. Both appraisal theorists and enactivists stress the idea that the phenomenology of emotion, so construed, is essentially linked to our possibilities of action given a relation with the environment. In a broad understanding of action tendencies, we can describe these relations as possible behavioral outcomes an organism may experience in a given moment. Again, phenomenology is described as part of other patterns already considered, casting doubts on the decision to separate it into its own category.

As a result, the individuation of phenomenological patterns as a separate category of patterns that would be candidates for correspondence and coordination seems unpromising. Either we get stuck with problems in grounding concepts in first-person experience, hence precluding us from establishing any claims regarding their variability, or, if we can overcome such an obstacle, we would be able to reduce phenomenal patterns to other patterns which turn out to be the relevant ones to decide for or against variability. Consequently, I propose leaving the qualitative character of emotional experience separate from the problem of variability or taking it as a result of other relevant patterns.

2.3 How to understand variability?

Having discussed the problems emotion researchers face when attempting to individuate patterns of responses in order to determine whether coordination and correspondence hold, we come to two central questions:

- How should we understand VT?
- According to this account of VT, is this thesis well-established?

Regarding the first question, recall the formulation of VT as presented at the beginning of this chapter:

Variability Thesis (VT) Emotions are naturally disjoined phenomena.

According to the preceding discussion, VT is analyzed in terms of a disjunction between the following two claims:

No one-to-one correspondence thesis (NOC) There is no one-to-one correspondence between emotion categories and neural, physiological, behavioral, expressive, or phenomenological patterns of responses.

Low coordination thesis (LC*) Variables in the neural, physiological, behavioral, expressive, and phenomenological domains do not constitute well-defined patterns of responses (i.e., display low correlations and do not constitute a well-defined set of causal relations either among variables in a domain or between variables in different domains).

Each of these claims, in turn, demanded analysis. There are two initial questions to answer in order to carry out such an analysis. First, which patterns of responses are relevant to distinguish emotions from one another? In other words, which are the patterns on which variability is to be evaluated? Second, once we have determined which types of patterns are relevant to decide questions about variability, we can raise the question of how to individuate those patterns in an empirically tractable manner.

In my view, the discussion in this chapter suggests that only three types of patterns or domains are relevant for questions about variability. These are neural, physiological, and behavioral patterns. The reasons why I exclude expressive and phenomenological patterns have already been presented and can be synthesized as follows. For both types of patterns, either they are relevant as subsets of behavioral responses, in which case they reduce to the behavioral domain, or they do not tell much about distinctions between emotions and thus about the issue of how emotions form kinds. In the expressive case, I argued that universality of expression does not imply homogeneity in terms of emotion categories, and at best it is only interesting as part of how we individuate and attribute emotions folk-psychologically. In the case of phenomenology, I argued that either it is cashed out in scientifically intractable terms (e.g. qualitative experience) or, if understood in other terms, reduces to neural, physiological, or behavioral responses. Hence, we can exclude these domains from the discussion regarding variability.

These arguments suggest the following account of VT as the disjunction of:

No one-to-one correspondence thesis (NOC') There is no one-to-one correspondence between emotion categories and neural, physiological, and behavioral, patterns of responses.

Low coordination thesis (LC') Variables in the neural, physiological, and behavioral domains do not constitute well-defined patterns of responses.

Having delimited the set of relevant domains to the neural, physiological, and behavioral domains, we can now approach the second question: how should we individuate these patterns in order to empirically evaluate VT? Offering a definite answer to this question requires a much deeper discussion that what I have and can offer here. Nevertheless, I offer the following working hypotheses.

In my view, evidence for neural patterns suggests that they are best individuated in terms of functional, distributed networks as found by multivariate analyses. Given the past failures to find anatomical regions or intrinsic networks corresponding to each emotion, the best candidate for neural patterns is functional locationism. Regarding physiological patterns, the question is much more complex. As I argued above, we should understand physiological patterns as sets of cardiac, electrodermal, and respiratory variables which are either causally related to one another or are robustly correlated so as to support distinctions between emotions.

Lastly, regarding behavioral patterns, our best candidate account is in terms of action tendencies, that is, dispositions to act in certain ways. In this case, the challenge is to avoid rendering empirical evidence trivial by tying these action tendencies to the meaning of emotion terms. As will become clear in later chapters, this can be done by explicating emotion concepts into functional kinds in terms of a psychofunctionalist framework. On this view, behavioral disposition patterns are found by empirical investigation and constitute part of an emotion's functional description. This is the main claim I argue for in Part II.

Having adopted the aforementioned working individuation criteria, we can now approach the issue of whether VT is empirically well-supported. In my view, the empirical case for VT is rather strong. Regarding coordination, we observe a relative degree of coordination in the neural case in terms of functional networks, but evidence in the physiological and behavioral cases is not as promising. In the physiological domain, even on multivariate analyses, there is an important degree of overlap between different emotion categories in terms of their associated autonomic responses. It is also unclear whether there are coordinated packages of responses in terms of causally related or correlated variables. At best, these hypotheses have not been tested with these concepts in hand, precluding establishing the presence of coordinated patterns.

Lastly, regarding behavioral patterns, evidence is sparse given that most of empirical research on emotions has attempted to map them onto neural and physiological patterns. Yet, some claims are suggested by available findings. On one hand, it is unclear whether expressive patterns, understood as a part of behavioral outcomes, are uniform enough to support a coordination claim for the behavioral domain. This is because both methodologically and empirically universality is controversial. On the other hand, even though evidence on the phenomenology of emotions that can be reduced to behavioral outcomes (e.g. Davitz's dictionary) does suggest some ways to differentiate emotions in terms of the ensuing action tendencies, the grain on which we must carve behavioral responses is yet to be determined. Again, I shall offer an account that solves this problem below. Nonetheless, for the purposes of determining the case for or against VT, I believe we must remain at least agnostic regarding the coordination of behavioral outcomes before more empirical evidence can be gathered.

Consequently, overall, VT seems to be relatively well-supported. If this is true, then, it seems that the Empirical Challenge has not been overcome yet. To be clear, let us apply the preceding analyses to reformulate this challenge. At the beginning of this chapter, I presented the Empirical Challenge as follows:

Empirical Challenge Provide a scientifically meaningful theoretical framework that establishes correspondences between emotion categories and well-coordinated patterns of neural, physiological, expressive, behavioral, and phenomenological responses.

This definition already captures the idea that VT should be understood as the disjunction of a correspondence and a coordination claim. This can be seen in that it spells out a rejection of VT in terms of a conjunction of $\neg NOC$ and $\neg LC$. Nevertheless, I have suggested that we do not include expressive or phenomenological patterns in this definition. As a result, the Empirical Challenge should be formulated in the following terms:

Empirical Challenge Provide a scientifically meaningful theoretical framework that establishes correspondences between emotion categories and well-coordinated patterns of neural, physiological, and behavioral responses.

As I have suggested, it seems that empirical evidence does not enable us to reject VT, making the Empirical Challenge a pressing obstacle for a unified scientific theory of emotions. This leads to a dilemma. Either we accept that emotions are disjoined phenomena, leading to taking emotion categories as arbitrary, which may lead to eliminativism, or we offer an account of emotion categories that surpasses the Empirical Challenge in spite of the empirical evidence in its support. In Part II, I will sketch meta-theoretical criteria for such an account, rendering the VT unproblematic and theoretically tractable. Before I discuss this solution though, let us consider the Empirical Challenge together with the Theoretical Challenge and explore their consequences.

Chapter 3

Interlude

Can we study emotions scientifically?

In the previous chapters, I have presented what I take to be the two main challenges for a scientific theory of emotions. These are what I have called the Theoretical Challenge and the Empirical Challenge. The Theoretical Challenge states that a satisfactory theory of emotions is one that explains the phenomena covered by the vernacular term "emotion" with a common set of explanatory resources. The Empirical Challenge claims that we an empirically tractable theory of emotions is one that allows us to reject the Variability Thesis, that is, one that enables us to find correspondences between emotion categories and well-defined (coordinated) neural, physiological, and behavioral patterns of responses.

First, on the Theoretical Challenge, following Griffith's (1997) argumentative strategy, I claimed that none of the best scientific theories of emotions succeed in offering a solution. These are basic emotion theories, appraisal theories, and psychological constructionist theories. All of these theories run into difficulties either internal to the theories' conceptual resources (e.g., problems with 'basicality' for BET or problems with 'core affect' for constructionism), their structure as scientific theories (e.g., involve unwarranted ad-hoc moves and *ceteris paribus* clauses), or successfully explaining the phenomenon at hand (e.g., making emotion categories arbitrary or not explaining all emotions). Hence, Griffiths's analysis, at least in part, still stands to this day.

Second, concerning the Empirical Challenge, I argued that it should be understood in terms of finding correspondences between emotion categories and well-coordinated patters of neural, physiological, and behavioral responses. For each of these domains, I argued that scientists must be clearer about what counts as a pattern candidate for correspondence. In the case of neural patterns, the best alternative is to seek for functional networks in the brain; for physiological patterns, we must be clear about causal connections between different variables as well as have a clear account of correlations between them; and for behavioral patterns, I argued that they should be cashed out in terms of action tendencies which are to be empirically investigated and not posited as a matter of definition.

In this interlude, I want to devote some time to some final remarks regarding these challenges. If we are to meet them, we must be clear about their aims, their demands, and their foundations. To do this, I first explain some points in common and some divergences. This will provide an idea of what would count as an answer to one or another, and whether it is possible to attack both challenges at once. Second, I discuss some consequences of past failures to meet the challenges. I explore two alternative replies to the current state of the art, namely, eliminativism and revisionism. In my view, we must opt for a revisionist solution. Thus, I argue against eliminativism and provide grounds to motivate revisionism. Lastly, I raise some questions about how to approach these challenges that will give some structure to the second part of our discussion.

3.1 The Two Challenges Considered

3.1.1 Where the challenges meet

As I explained above, the Theoretical Challenge requires an account of vernacular emotion categories. In other words, a theory that successfully meets this challenge is one that takes the differences of folk emotion categories into consideration. If this is correct, we can raise the question: does the same hold for the Empirical Challenge?

At a first glance, it seems that the Empirical Challenge does not require accounting for vernacular concepts of emotions. As Scarantino (2012) has argued, science could find specific and consistent mechanisms underlying emotions that do not map one-to-one onto vernacular emotion concepts. For example, we may find that fear corresponds to three kinds of mechanisms, anger to two kinds of mechanisms, and so on. It would thus appear that responding to the Empirical Challenge does not demand considering folk-psychological kinds at all.

Yet, there is a difficulty I will explain later in chapter 5. In my view, marking a strong division between folk-psychological concepts and scientific concepts runs the risk of changing the subject. When discussing dimensional theories, I claimed that one problem they run into the problem of making emotion taxonomies arbitrary and thus not accounting for the phenomena we call "emotions." The same could apply to the Empirical Challenge. Whatever specific and consistent mechanisms we may find, we must have clear criteria to call them mechanisms *of emotions*. Thus, I suggest, responding to the Empirical Challenge might also require considering vernacular concepts. How this can be done in detail will be a question I will discuss later.

This, however, seems to allow a response to the Empirical Challenge without answering the Theoretical Challenge. All that scientists require to tackle to the Empirical Challenge is some criteria, as vague as they can be, to call a mechanisms a mechanism of emotion. Consider LeDoux's experiments on fear conditioning in rats (Phelps & LeDoux, 2005). These experiments require some way of telling when a rat feels fear and has been thus conditioned successfully. Presumably, this occurs when the rat freezes after hearing the conditioned stimulus. On the face of it, all that the experi-

menter requires is this vague criterion to attribute fear to the rat: if the rat freezes, it is afraid. This seems far from a full-fledged commitment to a specific theory of emotions. If this is true, it follows that the Empirical Challenge is largely independent from the Theoretical Challenge.[30]

In my view, there three ways in which we can resist or at least qualify this objection. First, it is unclear whether this procedure is independent of theoretical commitments. The decision of studying fear in rats and the sense in which the scientist attributes fear to the rat already implies certain degree of theoretical commitment. Even if in the context of discovery this commitment is not explicit, in the context of justification it requires some background assumptions about why fear is a phenomenon that we can study by using animal models. In LeDoux's specific case, there is even an assumption that fear involves activation of the amygdala, hence the hypothesis that by lesioning this area, fear conditioning would be altered.

The fact that LeDoux's case involves this commitment also serves to show that once candidate theories have been proposed, the Empirical Challenge is no longer independent from those theories. Before we introduce a theory, it is plausible that scientists can start their investigation with very vague, perhaps folk characterizations of the phenomena. But once theories have been introduced, these theories shape what qualifies as an underlying mechanism of an emotion and hence what are the plausible hypotheses to test. If this is true, then given that theories are already on the table, the Empirical Challenge is no longer fully independent of the Theoretical Challenge. Finding mechanisms for emotions requires an analysis of how theories cash out these mechanisms and formulate their hypotheses.

Lastly, as I have formulated it, the Empirical Challenge stems from Barrett's discussion of basic emotion theories and discrete appraisal theories. This is the reason why the challenge asks for specific and consistent mechanisms, an assumption that arguably lies at the base of traditional accounts of emotions. This requirement does reveal an explicit theoretical commitment. If we formulate theories that do not expect specific and consistent mechanisms (a strategy that I intend to defend below), then the Empirical Challenge, at least in this version, is dissolved. As a result, the Empirical Challenge does call for a response to the Theoretical Challenge, at least as the debate stands currently, and at least in a very broad sense.

3.1.2 Where the challenges diverge

In spite of their common spirit, the two challenges diverge in important respects. Put differently, even though these challenges are connected in the sense that the Theoretical Challenge shapes the hypotheses that lead to the Empirical Challenge, they diverge in a much more nuanced sense. The Theoretical Challenge, as stated above, calls for a common framework to study all emotions. As such, replying to the challenge requires an overarching theory of emotions in general. Griffiths formulates the challenge in terms of kinds. If we take this formulation seriously, then the challenge becomes

[30] I thank Diana Pérez for raising this issue.

offering a theory that presents emotion as a single kind. In other words, overcoming the challenge implies showing how emotions form a class which allows generalizations and inductive inferences. I will leave the question of how this kind should be structured for the second part.

In contrast to this appeal to a general kind, the Empirical Challenge does not necessarily call for a general framework to study emotions. Instead, this challenge asks for an account of how specific emotions form kinds. This makes the Empirical Challenge less theoretically demanding than the Theoretical Challenge.

To make this clear, we can interpret the Empirical Challenge in two ways. On a weak interpretation, the challenge does not require all emotions to correspond to specific and consistent patterns of responses. All that is required is that at least some emotions have such correspondences. In this sense, views such as BET can meet the Empirical Challenge if they show that at least the so-called basic emotions correspond to specific and consistent mechanisms. I will call this the Weak Empirical Challenge.

On a stronger interpretation, however, the challenge demands all emotions to correspond to specific and consistent patterns. Even though this interpretation is more demanding than its weaker counterpart, it is still less stringent than the Theoretical Challenge. A possible response to this version of the challenge–the Strong Empirical Challenge, as it were–could account for all emotions as kinds but without committing to the idea that all emotions form the *same type* of kind. In other words, we could meet the Strong Empirical Challenge by showing that basic emotions form a particular type of kind while higher cognitive emotions form a different type of kind. This would make it possible to meet the Strong Empirical Challenge while failing to meet the Theoretical challenge.

If these interpretations are correct, it follows that meeting the Empirical Challenge does not require an answer of the same scope as an answer to the Theoretical Challenge. We can answer the Empirical Challenge by proposing different theories for each type of emotions, thus failing to offer an overarching theoretical framework. This is consistent with the fact discussed above that answers to the Theoretical Challenge inform the demands of the Empirical Challenge. In this case, it is as if we offered different replies to the Theoretical Challenge, failing to meet it as it stands in the literature but providing an account of what mechanisms are to be expected for each type of emotions.

How about the converse? Is it possible to meet the Theoretical Challenge without meeting the Empirical Challenge? At a first glance, we could offer a general theory of emotions that presents them as a higher order kind without committing to the claim that emotions in particular form kinds. Consider an analogy with other kinds such as gold. Even if gold forms a kind in terms of all of its elements sharing an essence, this does not mean that different types of gold form kinds themselves. There could be arbitrary distinctions between different types of gold such that there are no lower-level kinds relative to the general kind of gold. Similarly, it is presumably possible to offer a theory of emotions that groups them all into a single kind without particular emotions forming kinds themselves.

The preceding discussion raises a problem for this line of argument though. One of the criteria I have proposed to meet the Theoretical Challenge was to account for the vernacular distinctions between emotion categories. I criticized dimensional views on the grounds that they do not shed light on these distinctions, making classifications arbitrary. If this constitutes a failure to meet the Theoretical Challenge, it follows that meeting the challenge demands an account of the discreteness of emotion categories. Consequently, a response to the Theoretical Challenge must include an account of how particular emotion categories form kinds themselves. If this is correct, meeting the Theoretical Challenge would seem to imply a response to the Empirical Challenge, even in its strong form.

Nevertheless, I believe we can resist this conclusion and make a case for the possibility of meeting the Theoretical Challenge while leaving the Empirical Challenge aside. As I explained above, the Empirical Challenge not only appeals to the presence of a kind, but to a particular type of kind in terms of specific and consistent mechanisms. If we offer a theory of emotions that cashes them out in terms that do not demand specific and consistent mechanisms, then we would have met the Theoretical Challenge while dissolving the Empirical Challenge. On such an account, there would be no need to find specificity and consistency, thus rendering the Empirical Challenge uninteresting. This is the sort of account I will defend in Part II. Before discussing my proposal in detail, nonetheless, let us explore the consequences of failing to meet these challenges.

3.2 Failing to meet the challenge: Should we just give up?

What happens if we cannot find a theoretical framework that explains emotions and if we find no sense in which emotions correspond to well-coordinated packages of responses? Does this imply that a science of emotion is impossible in principle?

One pessimistic outcome is eliminativism. If there is no way to formulate a theory of emotions that is scientifically tractable, it follows that we should eliminate emotion categories from scientific discourse. Depending on whether the problem stems from general observations about the nature of scientific theories or whether it stems from problems with emotion concepts in particular, eliminativism may arise in two forms. I will call these Wide Eliminativism and Narrow Eliminativism respectively.

Wide Eliminativism is the claim that a theory of emotions is impossible in principle, not because of problems specific to emotion concepts but because of problems with a more general class of concepts. In its most influential form, Wide Eliminativism gets its voice in Churchland's (1981) argument. Churchland claims that there is no interesting sense in which we can study mental phenomena scientifically. The reason for this is that mental concepts, which have their roots in folk-psychological categories, have already failed to meet scientific standards. Baker (1993) spells out Churchland's argument (she calls it the *Argument from Science*) in the following terms:

(P1) Propositional-attitude concepts genuinely apply to humans if and only if they are underwritten by the best science of the mind.

(P2) The best science of the mind will not underwrite propositional-attitude concepts.
Therefore,

(C) Propositional-attitude concepts do not genuinely apply to humans. (Baker, 1993, p. 180)

Framed in this way, the argument concerns propositional-attitude concepts. Nevertheless, this argument is not what is at stake in the case of eliminativism regarding emotions. To see this, let us consider what I call narrow eliminativism, as proposed by Griffiths (1997).

As I explained in chapter 1, Griffiths argues that since none of our best theoretical frameworks capture emotions as a whole, then we must eliminate emotion categories from scientific discourse. The reasons he puts forward concern then emotions in particular, claiming nothing about other kinds of concepts in science. In order to evaluate Griffiths's argument, I updated his view to consider more recent theories of emotions. Moreover, I conceded that none of our best theories overcome the Theoretical Challenge. Hence, it would seem that Narrow Eliminativism is still on the table.

In contrast to Churchland, Griffiths does not think that there is no phenomenon corresponding to emotions. Instead, he believes that emotions correspond to a heterogeneous set of phenomena. This is different from Churchland's argument, which claims that there are no such things as folk-psychological states like belief or desires. Given this difference, I shall leave wide eliminativism aside, and focus on Griffith's narrow eliminativism.

I believe there are good reasons to resist Griffiths's conclusion though. In other words, the fact that our best theories cannot capture the phenomena in question does not entail eliminativism, even in this constrained form. The only thing that follows from this argument is that currently, we have no satisfactory theoretical framework. The possibility of offering such a framework, however, remains open.

Consider a classic example defenders of eliminativism have used in the past: the caloric theory of heat. The theory's failure to capture the phenomenon of temperature did not mean that temperature was not an object of scientific study. Rather, it meant that another theory had to be proposed, one that would capture the phenomenon successfully. The switch from the caloric theory to the analysis of temperature in terms of mean kinetic molecular energy did just that. Consequently, past failures do not entail eliminativism.

A more charitable interpretation of Griffiths's view could reply that not only does he show that no theoretical framework explains all and only the phenomena we call emotions, but that the phenomena themselves are heterogeneous and thus escape explanation and description under a common theory. The narrow eliminativist could hold that Griffiths's showed that it is not just that theories have failed to offer such a framework, but that the theories he considered are well-confirmed theories that show that emotions themselves are not a unitary phenomenon, but that they are rather two (or three) different kinds of phenomena. If this is true, it follows that an overarching

theory of emotions is impossible in principle, since we know already that we need different theories for each kind involved.

I do not think this interpretation holds up. First, as I claimed above, Griffiths's argument is outdated. This not only means that an update is required, but that the theories he considered are no longer our best theories. Consequently, it is not the case that those theories are now considered well-confirmed, nor that they reveal some intrinsic fact about emotions that precludes offering a general theory. Second, even in its updated version, one could turn things around and argue that precisely because our best theories have failed to explain emotions, they are bad theories. In other words, instead of taking sides with the theories, we can use these findings as arguments to reject the theories themselves.

As an alternative to eliminating emotion categories from science, I suggest we revise our scientific categories of emotions. Without exhausting the option of Revisionism, we cannot claim that a theory of emotions is impossible in principle. As long as a revisionist strategy is still available, eliminativism cannot get off the ground. Consequently, the rest of this work deals with the question of how to carry out such revisionism. As we will see, not only is revisionism still a possible alternative, but a promising one. If I succeed in defending this revisionist project, then it follows that (1) we can overcome the Theoretical Challenge, and (2) once we have a new theory in place, we can proceed to evaluate the empirical demands of the theory in order to revisit the Empirical Challenge.

3.3 What do we need to theorize about emotions?

In my view, there are three important pieces to answer the question of how to construct a scientifically interesting theory of emotions. The first is to make clear what are the criteria for a classificatory scheme to be considered scientific. This problem lies at the heart of criticisms claiming that emotions do not form a natural kind, such as Griffiths's and Barrett's. By claiming that emotions do not form a natural kind, they suggest that emotion categories are not objects of scientific investigation. In order to respond to these arguments, we must make clear what role the notion of "natural kind" is supposed to play in these arguments and in considerations about scientific theories in general. This will be the central question I will discuss in chapter 4.

Besides having an idea of what counts as a scientifically respectable kind, the next step is to determine how we can construct scientifically interesting taxonomies given the criteria adopted in the previous step. To do this, we must first clarify what is it exactly that we intend to explain in our candidate vocabulary, that is, what the explanandum phenomenon for our theory is. In other words, we need to *reconstitute the phenomenon* (Bechtel & Richardson, 2000/2010). I shall discuss this issue in chapter 5.

Lastly, once we have an idea of how to fix the explanandum and what types of vocabulary are available to us to construct a scientific theory, we must apply these results to the case of emotions. We must ask what are the conditions under which we

would say that a kind corresponds to emotions as an explanandum phenomenon, and which vocabulary is best suited to construct a scientific theory of emotions. This will occupy the final chapter (chapter 6).

If we succeed in answering these questions, we will have offered an account of how to construct a scientifically interesting theory of emotions, this allowing a response to the Theoretical Challenge. Additionally, we will have made progress on how to attack the Empirical Challenge, since offering the theory will give us ways to formulate testable hypotheses and tools to interpret empirical evidence already gathered.

Part II

The Solution

Chapter 4

Scientific Kinds

There are numerous ways to classify objects in the world. Most objects we encounter in our everyday lives can be classified according to size, shape, color, etc. When we classify these objects, we label the class and assign it a name. Thus, we talk about big and small, round or square, red or blue objects, and so on. We can also specify what the name is meant to capture, what property is it that allows such groupings and that the name is intended to convey. For example, we can say that objects classified as round are those whose borders are at an equal distance from their center. We can further elaborate more precise descriptions of these conditions, using more refined languages that allow us to make better distinctions. This is the case of red objects, which we can classify as those whose surface reflects a certain range of wavelengths (making use of the vocabulary of physics), thus producing in us the sensation of red.

Not all classifications are interesting for scientific inquiry. Objects that lie 300 meters away from the Eiffel Tower do share a property in common (namely, lying 300 meters away from the Eiffel Tower), but this does not tell us much about what kinds of objects we will encounter. Objects that satisfy this property may include cars, people, animals, buildings, etc. In contrast, other classes seem to be scientifically meaningful. Once we discover that all molecules of water are composed of two atoms of hydrogen and one of oxygen, we can expect them to behave in certain ways, such as them boiling at 100°C at sea level or forming solid structures below 0°C. The question then is: what distinguishes classes that are scientifically interesting from those that are not?

It is common to phrase this question in terms of *natural kinds*: what defines the genuine natural kinds from the merely appearance of natural kindhood? Presumably, it is because genuine natural kinds "carve nature at its joints" that they are the proper objects of scientific inquiry.[31] But what does it mean for a kind to carve nature at its joints? And why is this important for scientific investigation?

[31] This view is frequently ascribed to Plato. In the *Phaedrus*, Socrates presents two forms of discourse which the interlocutors follow in their discussion on *eros*. One we can call composition or synthesis, and consists in subsuming different particulars under one idea. The second can be called division or analysis, and consists in dividing things into elements. When presenting analysis, Socrates says: "The second principle is that of division into species *according to the natural formation, where the joint is*, not breaking any part as a bad carver might" (Plato, n.d., 265e; my emphasis). Interestingly, Plato was not discussing the division of objects in the world, nor problems related to induction or scientific inquiry. Rather, he is suggesting different methods

In this chapter, I will defend a reformulation of the latter question. I will claim that questions regarding scientific kinds require a pluralistic view of kinds that takes some distance from the tradition of natural kinds. In other words, I will argue that there may be many answers the question of what makes some classes interesting for scientific inquiry. What determines these answers are the explanatory interests of a given discipline and the conceptual resources with which we characterize the object of study. If this is so, the question of what kind of scientific object are emotions must be approached relative to the disciplines that study it, rather than a metaphysics of natural kinds.

4.1 The tradition of natural kinds

4.1.1 Mill and the introduction of Kinds

Mill (1843/1974) is often credited as the father of the tradition of natural kinds. Nevertheless, as I will show below, much of what the tradition has inherited from Mill stems from conflating two different concepts in Mill's work. These are the concepts of 'real Kind' and of 'natural group.' Real Kinds are meant to capture the logic behind general terms, while natural groups are intended to explain the metaphysical and epistemological aspects underlying induction. By conflating these two concepts, the tradition has assumed a picture of natural kinds that does not accommodate actual scientific practice, and that overstates the role certain patterns of induction play in how we construct scientific categories. With this in mind, I will present Mill's ideas on kinds and induction, and connect them to the contemporary landscape of theories of natural kinds and what I believe are important drawbacks of the current debates.

In his investigation on induction, Mill suggested that there is an unlimited number of ways to construct classes, but only some tell us something about the world. He noticed that some classes are bound together only by the property connoted by their name. This would be the case of the class of 'Objects that lie 300 meters away from the Eiffel Tower'; the only shared property we can know of is the one that the name of the class makes reference to. Other classes, those such as water, have a number of properties beyond the one specified by the name; molecules of water are not only molecules of H_2O, but also form substances that boil beyond certain temperature or freeze below another.

Mill thought that classes of the second kind correspond to distinctions in nature. In his words:

> And if any one even chooses to say that the one classification is made by nature, the other by us for our convenience, he will be right; provided he means no more than this: Where a certain apparent difference between things (though perhaps in itself of little moment) answers to we know not

in which we can proceed in discourse and present our ideas. Yet, this phrase has been widely popularized in debates about natural kinds.

> what number of other differences, pervading not only their known properties, but properties yet undiscovered, it is not optional but imperative to recognise this difference as the foundation of a specific distinction; while, on the contrary, differences that are merely finite and determinate, like those designated by the words white, black, or red[32], may be disregarded if the purpose for which the classification is made does not require attention to those particular properties. (Mill, 1843/1974, p. 123)

Thus, Mill thought that classes for which we can discover further properties, those whose properties are unexhausted by their defining property and potentially infinite in number, are to be privileged and taken as mirroring distinctions in nature.[33] Moreover, he identified these classes with those that Aristotelians called *genera* or *species*. The differences between genera and species are not mere accidents, but are differences in *kind*. In turn, differences in kind in the Aristotelian sense were taken to be differences in *essence*. The class of objects that lie 300 meters away from the Eiffel Tower differ from objects that do not in an accidental way. There is nothing essential to those objects that makes them members of the class. In contrast, members of the class WATER differ from other objects in the world because they have a specific essence (namely, being molecules of H_2O) that is not present in other objects. Furthermore, if any other object possessed said essence, it would necessarily count as member of the class WATER. Consequently, the difference between objects that lie 300 meters away from the Eiffel Tower is an accidental difference, whereas the difference between objects that are water and those that are not is a difference in kind.

These discussions introduced the term *kind* into recent philosophical literature. The term was first meant to capture those classes on which scientific inquiry was grounded, those for which science could discover new facts and properties. Even if Mill is wrong about what characterizes these kinds (see Hacking, 1991), it is important to note what the term is meant to convey. Notice, however, that Mill did not yet introduce the adjective 'natural' to designate these kinds. He thought that these interesting type of kinds would mirror distinctions in nature, but did not qualify them as 'natural kinds.' This is important because, as I will show below, the adjective 'natural' has become a source of confusion among philosophers of science and metaphysicians alike.

Even though Mill is often attributed the introduction of the notion of 'natural kind,' he uses two different albeit related terms. The first is the term «real Kind». Mill introduces this term after the discussion above, which is a discussion on the logic behind some general terms and their classificatory purposes. Real Kinds are those to which we assign a general term as a name, and are among the kinds that constitute interesting objects of scientific investigation. These are kinds that are "distinguished

[32] Notice that Mill thinks that classes grouped by color are examples of uninteresting classes. I do not share this view, but I shall leave this discussion aside.

[33] It is important to clarify that Mill did not believe that classes exhausted by their defining property where not also classes carved by nature. He recognizes that both types of classes are natural in a sense, and that classification itself is a human activity. What he does believe is that only classes for which there is more to discover ground scientific inquiry.

from all other classes by an indeterminate multitude of properties not derivable from one another" (Mill, 1843/1974, p. 126).

Nevertheless, the notion of a «real Kind» must not be confused with a the second notion in Mill's work, namely, that of a «natural group».[34] In the chapter called "Of Classification, as Subsidiary to Induction", Mill revisits some of the ideas regarding real Kinds, although with a difference in focus. It is clear from the title that in this case, classification is discussed not in relation to the use of general names, but as part of our inductive practices. Mill even starts this chapter by making such distinction. Here, classification is not taken as the division of things according to the use of a name (thus it is not the logical structure of classification that is being discussed), but as a problem of "[how] To provide that things shall be thought of in such groups, and those groups in such an order, as will best conduce to the remembrance and to the ascertainment of their laws" (Mill, 1843/1974, p. 712).[35]

Mill repeats a point made earlier in his first discussion of classification: there are infinite ways of logically constructing classes. But, similarly to the previous discussion, there are special ways to construct classes that, in this case, serve the purposes of scientific inquiry. In his view, scientific classification works by grouping objects in order to make the most generalizations possible. These groupings will correspond, whenever possible, to the properties that cause other properties of the class. Yet, it is the most salient effect that will serve to diagnose the class, as we do not know the causes beforehand. Therefore, a group of objects will be said to be *natural* if it is constructed in virtue of its most general similarity. These similarities may not be obvious from the outset, and their discovery constitutes one of the aims of scientific inquiry.

Consider the case of water. Saying that instances of water form a natural group amounts to saying that they share a number of properties in common, some of which are caused by some others. At the beginning of the investigation, we do not know which are the properties that cause one another, and thus we identify the objects by its most salient effects: they are generally found as transparent liquids, we are able to drink them, etc. In this sense, we determine certain salient similarities among different instances and group them under a class accordingly.

However, this is not yet sufficient for us to ascertain that water forms a natural group. Besides a general degree of similarity, it is also necessary that we group things according to those similarities that allow us to individuate the kind as best as possible. In Mill's words:

> [...] when we are studying objects not for any special practical end, but for the sake of extending our knowledge of the whole of their properties and relations, we must consider as the most important attributes, those which contribute most, either by themselves or by their effects, to render things

[34] The distinction between real Kinds and natural groups is discussed in detail by Magnus (2015).

[35] Moreover, in the previous discussion on real Kinds, Mill claims at several points that distinctions between real Kinds are interesting for the logician, and that different Kinds constitute different *logical species*. See Mill (1843/1974, p.123).

> like one another, and unlike other things; which give the class composed of them the most marked individuality [...]. (Mill, 1843/1974, p. 716)

In this sense, not only similarity, but also individuality, constitutes a necessary condition for a natural group. To put it differently, a natural group must be a group of objects naturally similar to each other in important respects, such that these respects determine the criteria to judge whether something is a member of the group or not, thus allowing us to individuate the group.[36]

To take the case of water again, once we have identified the class superficially, we can then investigate further and discover that these objects share the property of being molecules of H_2O, that they boil at 100°C, and other related facts about water. By identifying the molecular composition as central to what water is, we are formulating a definition that allows us to summarize our body of knowledge about water, as it is due to this molecular composition that other properties follow. As a result, we have found a way of ascertaining certain laws or facts about water by using this scheme of classification. Furthermore, any object that shares this property will be counted as member of the class, and thus this property becomes the individuating feature of the group.

Insofar as the notion of «natural group» is invoked in the discussion of classification as a task of dividing objects in the world, the distinction is metaphysical and epistemological. It is metaphysical, in the sense of capturing something about how objects are organized independently of ourselves. And it is epistemological, because it intends to say something about the best way to classify things for the purposes of induction and scientific knowledge.

Having presented the notions of «real Kind» and «natural group» as suggested by Mill, we can now raise the following question: do all natural groups form real Kinds and viceversa? As explained above, the notion of «real Kind» is a logical one, one intended to specify the logic behind the use of some general terms. In contrast, the notion of «natural group» is metaphysical and epistemological, intended to capture the ways in which objects are grouped together in nature and to make clear what is the best way in which we can construct taxonomies to capture such groupings.

Mill claims that distinctions in terms of real Kinds correspond to distinctions in nature in the sense that real Kinds form classes that are subject to further scientific inquiry. According to Mill, the fact that two groups differ in an inexhaustible number of ways is a symptom that these groups differ in nature. In this sense, all real Kinds are natural groups. Yet, natural group distinctions are not exhausted by distinctions between real Kinds. The reason is that there may be natural groups that do not have an indefinite and inexhaustible number of difference, therefore not forming different real Kinds. Mill uses the examples of differences among species to claim that these may only have a select number of properties distinguishing them. For instance, raspberries, which are members of the genus *Rubus*, do not differ from other members of the genus,

[36] Mill does not allow for natural groups or Kinds to be fuzzy. See Mill (1843/1974, p. 720).

such as roses, in an inexhaustible set of properties. Yet, they are classified differently by botany, and as such form different natural groups.

In sum, Mill introduced the notion of «real Kind» to characterize the logic behind some general terms, and the notion of «natural group» to capture distinctions in nature that constitute the objects of scientific inquiry (and hence induction). Unfortunately, many philosophers in the debates about natural kinds have missed this distinction. As a consequence, they take Mill's real Kinds as the precursors of what we now call natural kinds. This gives rise to the idea that natural kinds are individuated by their essences, as real Kinds are for Mill. In other words, this confusion has lead to one of the most important accounts of natural kinds: essentialism.

4.1.2 Essentialism

Recall that the question about what makes certain kinds interesting for science amounts to asking what makes certain groupings available for inductive inferences. Mill thought that induction depends on certain groups being formed in nature itself in virtue of their members possessing a number of salient properties, some of which allow us individuate the group and extend our knowledge. Given that some of these groups were individuated by their essence, this gave rise to the idea that perhaps all natural groups were so individuated. This view, which Mill rightly traces back to Aristotle, is the backbone of *essentialism*.

Essentialism can be defined as the view that natural kinds are individuated by a specific, epistemically privileged property that, by itself, warrants membership to the kind and serves justify inductive inferences across a kind's members—an essence. Contemporary essentialism is attributed mainly to the work of Kripke (1972) and Putnam (1970/1977, 1975). Interestingly, neither Kripke nor Putnam offered a defense of essentialism itself. Rather, they presupposed essentialism to develop their causal view on reference. Their idea is that the meaning of many general terms such as 'water' or 'gold' was determined, not by their intension, but by their extension, and that these extensions correspond to natural kinds understood as groups individuated by an essence.[37]

To understand the role essentialism plays in Kripke's and Putnam's view, consider the proposition 'Water is H_2O.' This is an identity statement, connecting the terms 'water' and 'H_2O.' Presumably, if this is an identity statement, it must be necessarily true, that is, it must be true in every possible world. The problem is that the intensions associated with 'water' and 'H_2O' are not the same. 'Water' can be described as 'A transparent substance that living organisms drink to survive' (and other descriptions associated with water), whereas 'H_2O' is described as 'A molecule consisting in two atoms of hydrogen and one atom of oxygen.' Given that their intensions are different,

[37] On one construal, essences may be understood in a logical sense, as that which defines a kind. This is not the sense in which I will use the term essence, and I think this is not the sense intended by essentialists about kinds. In my view, essentialists appeal to a stronger construal of essences in terms of properties that objects possess independently of our categories. In this chapter, I use the term "essence" in this latter sense.

how is it possible that the identity statement connecting them is necessarily true? It could perhaps be the case that in some possible world, chemists did not discover that water is H_2O, and as such the identity statement would be false. For example, in Putnam's Twin Earth (Putnam, 1975), chemists assigned the name 'Water' to substance XYZ, a substance satisfying all superficial properties of H_2O but differing in its chemical composition. In Twin Earth, chemists would not consider 'Water is H_2O' to be true, but the statement 'Water is XYZ' instead.

What Kripke and Putnam argued is that the problem arises only if we endorse the view that intensions determine the extension of the terms involved. If we abandon such a view, we can see that the extension of 'Water' and 'H_2O' in our world is the same, and therefore the identity statement connecting them is true. Furthermore, given our assignment of the terms to this extension, we see that in every possible world, 'Water' (assigned to the extension we assign to it in our world) always refers to the substance which we refer to when using the name 'H_2O.' In the case of Twin Earth, Twin Earthians have assigned the name 'Water' to a different extension, and in this sense they use the name differently from us. When we take into account such assignment of names, we see that 'Water is H_2O' is true in every possible world, even if in other possible worlds the assignment has been different. As long as we understand that our use of names refers to a particular extension, and that this extension remains the same in every possible world in which it exists, then our identity statement would hold even in worlds where other names are used to refer to them.

This argument runs if we presuppose that the extensions associated with 'Water' and 'H_2O' remain constant in every possible world where they exist. In other words, we need to presuppose that in every world were water exists, it is H_2O. This is the essentialist intuition underlying Kripke's and Putnam's argument, the intuition that if water is indeed H_2O, then it is so *essentially*. By 'essentially', they mean that an object that looked like water but was not H_2O (such as XYZ) would not be water at all; hence, H_2O would be the *essence* of water.

Salmon (1981), in his discussion of Kripke's view, pointed out that there are two theses involved in the argument above. The first is the thesis that names such as 'Water,' i.e. natural kind terms, refer to classes united in virtue of some essence; this is what he calls the Semantic Thesis. The second is the thesis that there are in the world classes united in virtue of some essence; this is what he calls the Metaphysical Thesis. As can be seen from the above, the Metaphysical Thesis does not follow from the Semantic Thesis. It might be the case that we assign natural kind terms as if there were such essentially bounded classes, even if there were no such classes. Thus, essentialism is presupposed, rather than defended, by Kripke's (and Putnam's) theory of reference.

Essentialism, taken as the claim that natural kinds are groups in nature formed in virtue of an essence, does provide an answer to what makes certain kinds useful for induction though. When we make inductive inferences, we observe a limited set of phenomena and we project certain properties (or predicates) to other unobserved instances of that phenomenon. We would be right in projecting these properties if the

instances above shared an essence, since we would be rightly attributing properties to things that will at least probably have them. For example, if we examine some samples of water and determine that they boil at 100°C at sea level, then we would be right in inferring that for any other substance with the same chemical composition (H_2O), it would be the case that it will boil at 100°C at sea level.

Yet, even if essentialism offers a theory for some interesting scientific kinds, it fails when it comes to others. Consider the case of lilies discussed by Dupré (1981). Lilies are referred to in biological taxonomy as members of the family *Liliaceae*. These include paradigm cases of lilies such as the Lonely Lily (genus *Eremocrinum*) and the Desert Lily (genus *Hesperocallis*). Yet, the family of *Liliaceae* also includes other plants that, under the common usage of the term 'Lily,' would not be included. These are garlic and onion (genus *Allium*). If *Liliaceae* were to form a natural kind, then we would have to specify an essential property that all of its members share. However, its members are so different (as different as paradigmatic lilies and onions) that finding an essence seems unlikely. But even if there was such a property, what is intended with this way of classifying lilies is to capture, not a set of essential features uniting the class, but their common ancestry and place in evolutionary history. In other words, the scientific kind *Liliaceae* is meant to support inductive inferences in virtue of a common historical background, not a common essence.

This sort of objection has been raised widely throughout the literature, particularly by philosophers of biology (Griffiths, 1999; Hull, 1978). In general, the objection is that some sciences, in this case biology, do not classify their objects by virtue of their presumed essences, that is, in terms of a concrete property or microstructure shared by all members of the class. As we will see below, other sciences such as psychology and neuroscience also appear to classify in non-essentialist ways. If this is the case, then essentialism cannot offer a satisfactory account of scientific classification.[38] What we need, perhaps, is a wider account of classification that accounts for the kinds essentialism account for, but also includes classification in other sciences such as biology. These considerations lead to the traditional alternative to essentialism, Boyd's (1991; 1999a; 2010) *homeostatic-property-cluster* account.

4.1.3 The HPC account

In the second half of the twentieth century, Richard Boyd developed the *homeostatic-property-cluster account* (HPC account for short), designed to overcome some of the issues raised against essentialism. In contrast to essentialism, which was developed mostly due to concerns regarding metaphysics and the philosophy of language, the HPC account was developed in the context of philosophy of science. As such, the

[38] There are also a number of metaphysical problems with essentialism. These include problems with what counts as an essence, whether there may be disjunctive properties as essences, whether they require a number of modal features such as necessity, and the level at which essences can be located. Since my concern here is with accounts of kinds as explaining scientific classification rather than with their metaphysics, I will not go over these problems here. For an overview, however, see Khalidi (2013, Ch. 1)

HPC account emphasizes the actual practices of scientists in order to draw an account of how particular disciplines classify their objects.

Homeostatic Property Clusters

We have already established that members of a kind share a number of properties that allow us to succeed in our inductive inferences. According to the essentialist, one of these properties will have a special place in our epistemic practices, as it will constitute the essential property that warrants induction across members of a kind. In contrast to essentialism, the HPC account claims that for at least some natural kinds, it is not necessary that there is one property shared by all members of the kind (e.g. an essence). Instead, we see that many kinds in science merely *tend* to share properties, that is, they have a number of properties that tend to cluster together but may vary across members of the kind. This would be the case of the family *Liliaceae*, whose members tend to share but not always possess the same properties, such as being flowering plants with generally large and colorful flowers and air-borne fruits.

For the HPC account, the important question is not only which clustering of properties is associated with the kind, but *why* does this clustering obtain. In this view, a given clustering of properties is inductively interesting only in case it obtains for a reason. Specifically, the HPC account claims that there must be a mechanism that makes these properties tend to cluster together if the kind is to be considered a natural kind. Kinds are thus individuated, not by members possessing the same set of (essential) properties, but rather by the mechanisms underlying the homeostatic clustering of properties. Consider again the case of *Liliaceae*. Even if there are outliers such as onions and garlic, members of the kind have a number of properties that tend to cluster together, such as the size and color of their flowers. In this case, these properties cluster together because all members of the kind descend from a common ancestor. As such, mechanisms involved in reproduction and evolution warrant that the properties associated with the kind will tend to co-occur in a relatively stable fashion.

Following Boyd (1999), HPC kinds may be initially characterized as follows:

(1) There is a family of properties F that are contingently clustered in nature.

(2) The presence of some of the properties in F either favors the presence of other associated properties or is caused by underlying mechanisms that favor it (or both). This is what Boyd (sometimes metaphorically, sometimes literally) calls *homeostasis*. (See Boyd, 1999a, p. 143)

Under this account, then, natural kinds are homeostatic clusters of properties. But importantly, Boyd thinks that not just any clustering of properties counts. Only those clusterings which are causally relevant for the explanatory purposes of a discipline are acceptable as natural kinds. Hence, he adds another condition to his characterization:

(3) The clustering of properties in F has causal effects because of the conjoint occurrence of properties and the presence of their underlying mechanisms. (Boyd, 1999a, p. 143)

I will return to the justification behind this condition below, when we discuss to which extent do we need a metaphysics of kinds to offer an account of scientific classification. For now, it suffices to show that Boyd's HPC account is committed to a realist view of science in which natural kinds map onto the causal structure of the world (see particularly Boyd, 2010).

Boyd further adds other conditions to his account. These can be summarized as follows:

(4) There is a kind term t that is applied to things that have a homeostatic clustering of most properties in F.

(5) t has no analytic definition. The definition is given by the properties and mechanisms that underlie the kind. Hence, defining the kind is a matter of discovery and theoretical development.

(6) A member of the kind may display some but not all of the properties in F, and some but not all underlying mechanisms may be present (i.e. there may be *imperfect homeostasis*).

(7) Whenever there is imperfect homeostasis, the relevance of some properties in F or some mechanisms over others is an *a posteriori* theoretical issue, not an *a priori* conceptual one.

(8) There may (and will be) cases of extensional indeterminacy.

(9) The causal importance of F, together with the underlying homeostatic mechanisms, is such that the kind or property denoted by t is a *natural kind*.

(10) No refinement of t towards less extensionally vague clusters will preserve the naturalness of the kind. Either it will overstate explanatorily irrelevant distinctions, or lead to a neglect of explanatorily relevant similarities.

(11) The HPC that serves to define t is not individuated extensionally but historically. The properties determining membership to t may change over time without t changing its definition, depending on the causally or inductively relevant properties at a given time. (Boyd, 1999a, pp. 143-144)

Accommodation and discipline relativism

So far, kinds are individuated by the presence of a mechanism that causes a causally relevant clustering of properties. There are, however, a wide range of mechanisms that can be identified in nature, and therefore there would be a wide range of possible taxonomies. What makes it so that we prefer some taxonomy over another? In Boyd's account, the decision for a particular taxonomy depends on the explanatory demands of a particular discipline. For example, since biology is presumably in the business of explaining commonalities among organisms that have evolved from a common ancestor, reference to species and other clades as kinds is vital to its explanations. In this sense,

reference to species as a kind *accommodates* to the demands of biology. In contrast, other disciplines such as gardening would not find the class of *Liliaceae* useful for their purposes, as their interests might lie, for instance, in the aesthetic properties of plants. As a result, *Liliaceae* would not form a kind for gardening, as the clustering of properties in this class is no longer explanatorily relevant to the discipline.

This claim can be formulated as what Boyd calls the Accommodation Thesis:

Accommodation Thesis We are able to identify true generalizations because we accommodate our inductive practices to the causal structure of the world by using a vocabulary that is itself accommodated to relevant causal structures.

Let us unpack this thesis. Each discipline has a number inductive and inferential practices along with some conceptual resources to explain the phenomena its interested in. These constitute what Boyd calls a *disciplinary matrix*. In order for these practices to be successful, each disciplinary matrix M will demand certain fit between its concepts and classificatory schemes and the causal structures relevant to the discipline. These are M's *accommodation demands*. Given that questions about natural kinds are questions about our inductive practices, the problem of how to define natural kinds can be formulated as a one regarding the contribution of natural kind terms and concepts to the satisfaction of the accommodation demands of different disciplinary matrices.

Boyd distinguishes two senses in which we define natural kind terms. On the one hand, for a disciplinary matrix M, we can define the kind by the explanatory role that the use of a natural kind term t plays in satisfying M's accommodation demands. For example, we can define an element by its relation to other elements in the periodic table, or a species by its phylogenetic position. This is what Boyd calls a *programmatic definition*. On the other hand, we can define t by appealing to the properties associated with the kind. In this second sense, we define an element by appealing to properties such as its atomic number; in the case of the species, we may appeal to the derived characters that tie members of the species together. Boyd calls these *explanatory definitions*.

An important feature of explanatory definitions is that the properties that figure in them must be those in virtue of which reference to the kind satisfies M's accommodation demands. In other words, they must be those properties that enable the use of natural kind term t to play the role specified in the term's programmatic definition. Hence, there is a close relation between the two types of definition. Programmatic definitions set the role reference to the kind plays in the inductive and explanatory practices of a discipline, while explanatory definitions pick out the properties by which this role is satisfied.

One consequence of this relation is that the definition of a kind term is relative to the inductive practices of a given discipline. Since explanatory definitions depend on programmatic definitions, and programmatic definitions depend on what is it that a discipline wants to explain, then the natural kind vocabulary we use in a scientific context will be discipline-relative. This discipline dependency has lead some to criticize

the HPC account, claiming it is dangerously close to conventionalism. The objection holds that the HPC account makes it so that natural kinds are decided by convention, and not by nature itself (see Craver, 2009, for a discussion). Boyd himself accepts that his account involves some degree of conventionalism, but he stresses that the decision to accept a kind is not merely conventional, but responds to the sorts of causal structures that constitute the object of study of a discipline. Again, arbitrary clusters of properties, even if present in high frequencies, do not form natural kinds if they are not useful to understand the causal relations within and between the kinds involved. The class of *Liliaceae*, for instance, is a natural kind on this account, because it helps us understand the causal history that lead paradigmatic lilies, onions, and garlic to share the same cluster of properties. In contrast, the set of objects that lie 300 meters from the Eiffel Tower, even if they have a range of similar properties (e.g. being spatiotemporally located in France) do not engage in interesting causal relations, and therefore the class does not accommodate to the explanatory demands of any discipline.

Shortcomings of the HPC account

Despite its initial plausibility, the HPC account is not without shortcomings. Generally speaking, the main problem with the HPC account is that even though it aspires to offer a general account of natural kinds, it fails to capture many kinds in science. One such objection appeals to fundamental particles in physics (Chakravartty, 2007; Magnus, 2014). The argument is premised on the idea that if fundamental particles were to count as a natural kind on the HPC account, there would have to be an underlying homeostatic mechanism producing the property cluster that defines the kind. The presence of an underlying mechanism would imply that the particles in question are not fundamental. Thus, either they are not a natural kind or the HPC account is wrong. Since physics makes successful generalizations by making reference to fundamental particles as a kind, then the HPC account is wrong.

Similar lines of argument run for other types of kinds, including the kind that the HPC account was thought to be most suited for, species (Ereshefsky & Matthen, 2005; Ereshefsky & Reydon, 2015; Slater, 2015). According to this objection, the HPC account puts too much emphasis on similarities among the members of a putative kind, as it requires members of the kind to share more or less the same properties (that is what is intended with the notion of a property cluster). Biology, however, defines species in terms of their phylogeny, disregarding similarities almost completely. Hence, species, as classified in biology, fall outside the scope of HPC kinds.

Besides fundamental particles and species threatening the generality of the HPC account, there is yet another whole family of kinds that lies outside the HPC account's scope: multiply realized kinds. Multiply realized kinds are pervasive in psychology and neuroscience, and as such, they should be included in an account of what makes certain kinds scientifically interesting. To understand the problems concerning these kinds, consider the case of memory.

Neuroscientists distinguish between different types of memory. First, there is declarative and non-declarative (or procedural) memory. Declarative memory includes episodic and semantic memory, i.e., the recollection of past episodes or facts. Non-declarative memory involves cases of memory that do not involve the declaration of episodes or facts, such as learning how to ride a bicycle or conditioned learning. These two types of memory have been found to have different underlying mechanisms. Evidence for this claim comes particularly from lesion studies. This is the famous case of patient H.M., who lost parts of his hippocampus and other regions as part of a surgical procedure to decrease his epileptic seizures. After this lesion, H.M. retained many of his declarative memory abilities, such as recollection of his own biography. However, he lost the ability to form new memories and other abilities related to short term and procedural memory. This result invited distinguishing between declarative and non-declarative memory, and to examine the different mechanisms underlying each (Baars & Gage, 2010; Gazzaniga & Mangun, 2014).

On the HPC account, given the difference in mechanisms underlying different types of memory, there would be reason to think that memory is not a natural kind. This claim has been defended by Michaelian (2010), for instance. Michaelian thinks that we cannot give a unifying computational description of the various types of memory. If this is true, then we cannot unify these types of memory under a single cluster of properties, and thus we cannot make sense of memory as a natural kind. In a similar vein, Pöyhönen (2016) holds that memory fails to be a natural kind both under essentialist and under HPC accounts due to the fact that we have no unifying account of how much does memory extend, whether it should be considered only intracranially, transactionally, or as widely extended.

Consider a similar case, the case of psychiatric categories. Psychiatric categories are often individuated as collections of symptoms in different combinations. For instance, the DSM V diagnoses *major depressive disorder* by considering a list of nine symptoms, where any combination of five out of nine (in case they are not attributable to any other condition and in case they disrupt the patient's social or other kinds of functioning) suffices for the diagnosis. Importantly, episodes of major depressive disorder must not be attributable to the physiological effects of a particular substance or an underlying medical condition. In this sense, it is important that major depressive disorder has no identifiable underlying mechanism, but instead is instantiated in different ways across patients. What characterizes this disorder is not one mechanism causing a number of symptoms, but the co-occurrence and interaction of the symptoms themselves. As a result, given the relative disregard for a particular defining mechanism, it would appear as if psychiatric categories like depression were not members of the set of natural kinds.

This result is nevertheless puzzling. Major depressive disorder, as a category in the DSM and other diagnostic manuals (e.g. ICD-10) serves important epistemic purposes in psychiatry. First, it allows scientists to distinguish a number of patterns of behavior from other related patterns such as anxiety or bipolar disorders. Second, it has lead to a number of interesting hypotheses in neuroscience (e.g. the monoamine deficiency hypothesis). Lastly, it enables psychiatrists to identify important interactions between

major depressive disorder and other conditions such as schizophrenia or multiple sclerosis. As such, the category seems to serve the inductive practices of psychiatry, even if it does not have one unique underlying mechanism defining it. Rather, what seems to be accounting for the epistemic success of the category is that it unites several mechanisms under a common *functional pattern*. In other words, it is the functional unity of the category, that is, the fact that it unifies a constellation of symptoms under a common concept, that makes it useful for the purposes of psychiatry (see Cramer et al., 2016).

This functional character not unique to major depressive disorder. On the contrary, it is a common feature of psychiatric categories that they are functionally individuated. Godman (2013) discusses the case of *pathological withdrawal syndrome* (PWS) and its status as a natural kind. As she presents it, PWS is characterized by symptoms including apathy, depression, panic attacks, and some others. Interestingly, psychiatrists investigating PWS concluded that this syndrome is not an organic disorder, meaning that it is not individuated by some medical or physical illness. Yet, they are able to use the category to diagnose and successfully treat patients that suffer it. Patients even have similar experiences throughout the development of the syndrome, even at the stage of recovery. This suggests that PWS is an interesting psychiatric category whose use for scientific and therapeutic purposes does not depend on the identification of a particular underlying mechanism. Rather, it is individuated functionally and used successfully to make some generalizations about the condition.

The cases of memory and psychiatric categories, among others, show that there are kinds, including multiply realized kinds, whose use seems to depend on their functional unity rather than the identification of a homeostatic mechanism or an essence. Nevertheless, in the HPC account, we do not find a clear account of how these functional kinds do their job. The reason is that these kinds, insofar as they are instantiated in different systems and processes, fail to provide us clear reasons that support inductive inferences across members of the kind.

Millikan (1999) raised this issue when discussing Fodor's (1974) argument for the autonomy of psychology. Millikan examined kinds whose instances, according to Fodor, map onto instances of kinds at lower levels, but not always to members of the same lower level kind. For Fodor, these are the kinds present in psychology. Consider again the case of memory. Memory states map onto brain states (lower level kinds), but not to the same kind of brain states; declarative memory states map onto different mechanisms than those underlying non-declarative memory states. In Fodor's view, the multiple realizability of psychological kinds shows that psychology is an autonomous science insofar as its kinds are not reducible to the kinds of lower level sciences such as physics.

Millikan argues, against Fodor, that the sorts of kinds over which scientific generalizations can be made must be bound together either in virtue of their history (for which she coins the term «*historical natural kinds*», and which she associates with Boyd's HPC kinds (see Boyd (1999a))) or in virtue of some property members of the kind share ahistorically (for which she coins the term *«eternal natural kinds»*). The

reason behind this constraint is that for Millikan, there must be a good reason to think that unobserved members of a putative kind will possess at least some of the properties that observed instances do. In the case of historical natural kinds, there is some common history that led members of one kind to resemble one another in important respects. For example, we know that tigers we have not observed yet will most likely have stripes because they have evolved from the same ancestors as other tigers that do have stripes. On the other hand, eternal natural kinds are bound together by some eternal property that they have independently of spatiotemporal considerations. For example, if water is H_2O, then we can presumably claim that it has been H_2O as long as it has existed and as long and wherever it will exist.

The problem with multiply realized kinds, in Millikan's view, is that there is no good reason to expect that unobserved members will share the same properties as observed members of the putative kind. If multiply realized kinds are, by definition, realized in different instances of different lower level kinds, then the properties present in one instance of the higher level kind might not be present in another instance. Thus, laws in higher level sciences become plagued with exceptions, thus precluding any interesting inductive inference. In the case of psychology we can, at best, have generalizations localized to a particular species, but this is only because psychological kinds within a species are realized in relevantly similar ways, claims Millikan. For instance, it is plausible that human brains instantiate non-declarative memory states in similar ways, and thus we can project from one instance of non-declarative memory to another. Yet, when we jump from humans to, for instance, Martians, the instantiation of non-declarative memory might be so different that none of what we know about human non-declarative memory states will apply to Martian psychology.

If Millikan's argument is right, then psychological kinds, taken as kinds in a general science covering different species and systems, are not scientifically interesting, as they do not support the sort of generalizations that science is interested in making. Moreover, the result would be generalizable to all sorts of functional kinds, including psychological but also psychiatric kinds. Insofar as functional kinds are multiply realizable, they are neither historical kinds nor eternal kinds in Millikan's sense. Hence, either we conclude that functional kinds are not scientific kinds at all, or we reply to Millikan with an account of how functional kinds enable us to be successful in our epistemic practices and hence count as legitimate scientific kinds. Unfortunately, neither essentialism nor the HPC account provide good prospects. Given the multiple realizability of functional kinds, they do not constitute kinds tied together by some essence; and given that Millikan associates HPC kinds with historical kinds, functional kinds do not constitute HPC kinds either.

Boyd (1999b) thinks that some multiply realizable kinds may be natural kinds in spite of Millikan's observations. He invites us to consider livers. Livers are multiply realized insofar as there may be differences across human livers and between human livers and those of other species. Yet, "liver" is a term that functions as a natural kind in anatomy: it involves certain causally relevant properties and serves the accommodation demands of the discipline in which it functions. In Boyd's view, this sort of multiple

realization is unproblematic because of what he calls *replacement stability*. If I have surgery on my liver, some of its microstructural properties change but at least many of its macroscopical properties remain stable. The same holds across different human livers which in spite of their differences, share a number of macroscopical properties. This explains why "human liver" is a term which can be used to pick out a natural kind in anatomy: the properties associated with the kind remain stable across realizations.

If Boyd is right, then the same holds for many other multiply realizable kinds. This is why he thinks that it is possible to have multiply realizable natural kinds in case "the commonalities produced by the relevant replacement stabilizing processes are sufficiently robust and relevant to accommodation" (Boyd, 1999b, p. 95). This also invites a multi-level view of kinds. We can have human anatomy, mammalian anatomy, and even vertebrate anatomy. At higher levels of abstraction, kinds become less uniform but, as long as there is some stability across members of the kind, they can still figure in interesting scientific enterprises. If so, then the multiply realizable kinds of psychology may count as genuine natural kinds.

Reydon (2009) develops a similar argument to expand the HPC account to cover functional and multiply realizable kinds. In his view, traditionally accepted natural kinds served traditionally accepted explanations insofar as they provided the divisions on which explanation is supposed to work. By dividing things into natural kinds, we constrain our explanations to classes whose members share something that allows us to project these explanations throughout the class. This, he points out, is also what functional kinds do for mechanistic explanation.

A mechanistic explanation is a description of "how the component entities and activities are organized together such that [a] phenomenon occurs" (Craver, 2006, cited in Reydon, 2009, p. 730). The organization of these entities and how they produce the phenomenon is functional, insofar as it focuses on how the component entities relate to one another produce an output, i.e. to fulfill a certain function. Hence, in mechanistic explanation, mechanisms—and therefore the basis on which scientists will generalize—are individuated functionally. In Reydon's terms:

> What matters in [mechanistic explanations] is that particular parts perform functions under particular circumstances. The detailed ways in which these functions are actually realized are not important in the analysis of the overarching system [...]. Functionally defined kinds, then, serve as the "hinges" around which [mechanistic explanations] turn in the following sense. Reference to the functional kind to which a particular part of a system belongs is explanatory as the basis of a generalization about the behavior that it is expected to exhibit when placed in a particular environment. In addition, the existence of the various functional kinds is itself a phenomenon in need of an explanation, as it need to be explained how the black-boxed entities are able to realize the various functions that they realize as parts of systems and how these entities have come into existence in the first place. (Reydon, 2009, p. 731)

Reydon thinks that by given the role functional kinds play in mechanistic explanations, they should be regarded as natural kinds in the same sense as traditional examples do. Both traditional natural kinds and functional kinds act as the "hinges" that serve as the basis of certain generalizations. As a result, Reydon proposes an extension of the HPC account to include functional kinds. His proposal is to include the factors that cut things into kinds in particular types of explanation into the definition of an HPC kind. In Reydon's words, we can extend the HPC account as follows:

> (**HPC***). A particular natural kind term is defined by a combination of a particular Φ, *F**, and *H**, where
> Φ = the factor(s) that individuate(s) things as members of particular kinds in explanations (e.g., the capability to perform a particular causal role function),
> F* = the set of those particular properties that play central roles in the explanation of Φ and are found to repeatedly cluster in nature, where this clustering may be imperfect and exception-ridden,
> H* = the set of causal factors that underwrite this clustering. (Reydon, 2009, p. 734)

In this sense, if some scientific explanation (e.g. mechanistic explanation) carves kinds by appealing to the functional properties of some objects, these kinds will count as HPC kinds as well. Hence, the upshot of this proposal is, on one hand, accommodating functional kinds into the framework of the kinds captures by the HPC account, and on the other, showing a link between functional kinds and their role in mechanistic explanations.

These attempts to extend the HPC account to capture functional kinds are interesting on their own right. Yet, I believe that the problem regarding functional kinds, as well as many other problems concerning natural kinds, do not lie in the difficulties of accounting for them in terms of the HPC account or any other particular account. Rather, functional and multiply realizable kinds are problematic only if we hold on to the idea that all scientific kinds are useful for induction because of the same reason. Given that multiply realizable kinds differ in important respects from essentialist and HPC kinds, they turn out to be problematic, but only given the expectation that all scientific kinds are of one type. Once we are ready to recognize different types of scientific kinds and different reasons why scientists classify objects into kinds, can we see that functional and multiply realizable kinds are not problematic, but only different. In other words, by adopting a pluralistic view of scientific kinds, we can open up some space for functional kinds as well as other types of kinds.

4.2 Projectibility and scientific kinds

I began this chapter by asking the question of what sorts of kinds allow us to do science. I presented the traditional answer, which identifies the interesting kinds for science (i.e., scientific kinds) with natural kinds, and examined some of the theories of

natural kinds and their shortcomings. What the tradition missed, in my view, is that we need not identify scientific kinds with one type of natural kind. To see why this is so, let us go back a couple of steps.

Scientific kinds, in the most general sense, are kinds that allow us to do induction. In turn, doing induction is inferring something about an unobserved set of entities or events given already observed instances. This inference from the observed to the unobserved involves *projecting* properties from the former to the latter. If every sample of water we have observed boils at 100°C at sea level, then we can infer that future samples will be have in the same way and that they share the same properties that our observed samples have. Hence, we can project successfully.

Notice that when we project properties to unobserved instances, we project to other members of the *same kind*. We can project the property of boiling at 100°C to future samples of water because the form a kind with those that we have observed before. Thus, the question of what makes induction successful ties in together with the question of what determines kinds and how do we use kind terms in our epistemic practices. As Hacking (1994) puts it:

> Classification and generalization must be rejoined. To use a name for any kind is to be willing to make generalizations and from expectations about individuals of that kind. To use a common noun to classify is to use it, and to use it is to be willing to project it. (Hacking, 1994, p. 221)

The question of what makes scientific kinds interesting thus amounts to the question of what makes certain classifications inductively interesting for a scientific discipline, or what makes certain scientific classifications adequate for projection. This question is reminiscent of Goodman (1979/1983). Goodman saw the problem of induction as an application of the more general problem of projection, and therefore a solution to the latter would buy us a solution to the former. Moreover, by having a theory of projection we can also get closer to a theory of scientific kinds, as a theory of what makes it so that we can project certain properties or predicates and not others will shed light on what makes certain classifications useful for projecting.

Goodman formulated the problem of projection in terms of which hypotheses or predicates are worth adopting and how can we eliminate undesirable ones. He devised the famous 'grue' case to explain that if we limit ourselves to the relation between the evidence and the hypotheses that we can formulate based on it, we cannot reach a criterion to decide between desirable and undesirable hypotheses. The case can be presented as follows. Consider emeralds and the hypothesis "All emeralds are green." This hypothesis is supported by all our evidence so far, and therefore seems like a case where projection is not problematic. In other words, there seems to be no problem with projecting the predicate 'green' to all emeralds. However, we can also think of another predicate, 'grue,' that applies to all things examined before time t in case they are green, and to other things in case they are blue. Before time t, the hypothesis "All emeralds are grue" is equally supported by the evidence, and therefore, if we are willing to adopt the hypothesis "All emeralds are green," we should also be willing

to adopt the hypothesis "All emeralds are grue," even though these two hypotheses are inconsistent with each other. Importantly, evidence will not provide a criterion to decide between one of these two, as all evidence at time t supports both hypotheses. Thus, we need to look elsewhere for a criterion.

The solution proposed by Goodman appeals to the use of a predicate in past projections. The intuition underlying this solution is that the predicate 'green' has a better record in past successful projections, and therefore it is better *entrenched* into our body of knowledge. An entrenched predicate, in Goodman's account, is one that has been used successfully in the past and that is coextensive with other projectible predicates. Consequently, since we have used 'green' in the past instead of 'grue', we prefer 'green' and the hypotheses in which it figures, even if 'grue' is equally supported by the evidence.

This conclusion sheds light on why should we prefer some classifications over others, and thus why we use certain kinds for science. Goodman himself remarked that using entrenched predicates to group things entails adopting some classification scheme over others. In other words, we take some classes to be useful for science because they are well entrenched. Therefore, scientific kinds are, at the very least, entrenched kinds.

Quine (1969/1977), following Goodman, applied this solution to the case of natural kinds. He thought that Goodman's solution captured the intuition that green objects, being an entrenched part of our classificatory schemes, are more similar than grue objects in some important respect. In his words:

> [...] why do we expect the next [emerald] to be green rather than grue? The intuitive answer lies in similarity, however subjective. Two green emeralds are more similar than two grue ones would be if only one of the grue ones were green. Green things, or at least green emeralds, are a kind. A projectible predicate is one that is true of all and only the things of a kind. (Quine, 1969/1977, p. 157)

To put it differently, taking 'green' as a better entrenched predicate than 'grue' implies that in our classification, green objects will form a kind. Furthermore, due to the entrenchment of 'green,' we construct theories in which the property of greenhood is more important that the property of grueness. In Mill's terms, green is a salient property that allows us to individuate the kind, and therefore it is suited to figure into the ways in which we divide objects in the world.

This invites the view that all there is to scientific kinds is projectibility and entrenchment. Naturally, there may be resistance to this view. In particular, this view seems incompatible with realism about science. After all, if we hold to the view that kinds are projectible and entrenched classes, this makes scientific classification dependent on our interests and minds rather than on reality. For some, this may be an overly pragmatist view of kinds. Before advancing the argument, then, let us discuss the question of realism about kinds, and examine the consequences of holding a projectibility-view instead.

4.2.1 Realism and kinds

One of the motivations behind theorizing about natural kinds is the idea that science aims at the discovery of classes that exist in Reality (with a capital R), independently of our knowledge and interests. Natural kinds, presumably, capture Real groupings in nature, rather than our own arbitrary and contingent, empoverished ways of classifying objects in the world. Yet, I have claimed that all there is to natural kindhood is projectibility and entrenchment. This makes natural kinds dependent on our own epistemic stances, rather than in Reality. Therefore, the view I am advancing here seems to run counter to realism about science.

In this section, I will claim that the question of realism is a red herring. What is important at the moment is how our scientific classifications serve our epistemic purposes, rather than the question of whether we can say that some classifications are more privileged than others in the sense that they get closer to something we can call 'real.' This is not to say that one cannot be a realist about kinds though. What I want to claim is that questions about realism are of a different order than questions about how science constructs kinds. In other words, I want to claim that a pluralistic account of kinds of the sort I am proposing is compatible with either an endorsement or rejection of realism.

Both the essentialist account and the HPC account are committed to the idea that natural kinds ought to be real in some special sense. For the first, essences are special properties that exist in reality and that determine the ways in which objects will group themselves. And for the second, it is the real causal relations that lead to the clustering of properties that determine natural kindhood. In any case, both views and therefore much of the tradition has been committed to a realist account of kinds.

Why should we commit ourselves to such an account? One of the main motivations for realism in general comes from the no-miracles argument. According to this argument, we cannot understand scientific progress unless we accept that science approximates reality as it develops new theories and explanations. Otherwise, scientific progress becomes a miracle, mere luck in our wild guesses about the world. Applied to kinds, the argument holds that unless our scientific theories approximate the ultimate true taxonomy of reality, we cannot explain how is it that we are successful in projecting across certain classes of objects.

Now, does it really matter whether kinds approximate something like an ultimate taxonomy? Even though the tradition of natural kinds has been framed in the realist picture, the question that they are trying to ask in relation to scientific kinds is not one about reality but about *epistemic success*. Recall again the question with which I started this chapter: why are some classes better than others for the purposes of scientific induction? To answer this question, we need not say much about the problem of realism.

For the sake of argument, let us suppose first that realism about science is false and that science does not approximate reality. Yet, there are still some classificatory schemes that are better suited for our theories and explanations, or at least that we hold

to better standards than others for certain epistemic purposes. Just as the classification of lilies, garlics and onions in terms of the family *Liliaceae* was better suited for biology but not for gardening, there are classifications that accommodate (to use Boyd's term) better than others independently of whether science approximates reality. The question is thus why are some classifications better suited to some disciplines and therefore more epistemically successful than others. Claiming that it is because they are 'more real' does not explain much, as even if they were not, scientists would still hold some classifications as preferable over others. In other words, explaining the success of certain classifications in terms of them being real amounts to little more than foot-stamping realism, as Fine (1984) puts it.

Let us now suppose the contrary, that realism about science is true. Even if we suppose that realism is true, we have not said anything about the reasons why we should adopt a classificatory scheme over another. There are a myriad of ways to create classifications. What the Realist is asking is to find the one that 'best' reflects Reality. This presupposes either that there is only one privileged taxonomy that is the one that captures Reality the best, or that there are several taxonomies that can do the job. Limiting kinds to only one taxonomy over all is much too stringent and does not accommodate adequately to actual scientific practice. I will expand on this idea below. On the other hand, once we recognize that science works with a number of taxonomies, we would then need to ask not just which one is it that reflects Reality, but which one does it *best*. Yet, asking which one is 'best' is always asking which one is best *for something*. And since science is in the business of explaining and predicting (besides reflecting Reality if realism is true), it follows that we must then ask which taxonomy works best for these epistemic purposes. Thus, even if Realism is true, questions about what characterizes scientific kinds amount to questions about scientific projectibility and entrenchment.

This is not to say that we should opt for an anti-realist view of kinds. All I want to claim is that questions about how science picks kinds are independent from questions about realism. Asking whether scientific kinds are real is a whole different problem, one that we need not answer here. Instead, the central question for a theory of scientific kinds is one about the criteria by which a scientific discipline holds certain classes as projectible and well-entrenched.

Another issue related to realism about kinds is the idea that there must be one epistemically privileged taxonomy over all others. According to this view, there is one ideal taxonomy that would in principle capture the kinds that exist in nature. One could, additionally, call this taxonomy the 'real' one, as it were. This is what Hacking (1991) calls the Uniqueness Principle:

> There is a unique taxonomy in terms of natural kinds, that represents nature as it is, and reflects the network of causal laws. We do not have nor could we have a final taxonomy of anything, but any objective classification is right or wrong according as it captures part of the structure of the one true taxonomy of the universe. (Hacking, 1991, p. 111)

Hacking himself rejects this principle, arguing that the idea of a complete taxonomy is nonsensical. He thinks that we should instead develop a historical understanding of natural kinds as they figure in different disciplines and historical contexts. In this view, kinds are epistemic tools with which we do things, namely, with which we understand the world. As such, they may be changed and tailored according to our current epistemic commitments.

Similar arguments abound in the literature. Dupré (1981), for instance, presents it as the thesis of *promiscuous realism*. For Dupré, there are a myriad of similarity relations that support projectibility, and none of these relations should be taken as privileged in any sense. As a result, no one taxonomy gets the title of being more real than another, and therefore all taxonomies are equally real in some sense. After all, he claims, all taxonomies capture some range of similarity relations that may be relevant for a particular purpose.

Khalidi (2013) agrees with Dupré in this respect. For Khalidi, there is no unique best system of classification, since the choice of a particular systems depends on our purposes. However, he raises the issue of whether there are some purposes that we hold to better standards. In his case, he privileges epistemic purposes, understood as the discovery of real divisions in nature. Other purposes, such as aesthetic purposes (like classifying plants for gardening) do not map onto the causal network of the world, and therefore do not capture a real taxonomy. Ultimately, he claims, there is no one privileged taxonomy but there is a privileged class of taxonomies instead.

The argument I am advancing comes closer to Dupré's than to Khalidi's. I agree with both, as well as with Hacking, that the idea of a unique taxonomy, privileged over all others, is chimerical. If what matters about kinds is projectibility, and projectibility can obtain in many different ways depending on what predicates and hypotheses we take to be well-entrenched, then there is no single taxonomy to rule them all. There are, instead, many taxonomies that shift focus from some properties to others depending on what properties we take as relevant for our explanations.

However, I do not think that the taxonomies that are tailored according to these epistemic purposes are privileged in capturing a hidden reality. In this respect, I distance myself from Khalidi's view that taxonomies in terms of natural or scientific kinds must reflect a causal network present in the world. Again, adding the label of 'Real' to a set of taxonomies does not tell us why we adopt those taxonomies in the first place. Furthermore, Khalidi's view is premised on the idea that what underwrites scientific kinds are causal relations, but this seems to fall back to the idea that there must be a unique theory of natural kinds, one unique reason why all successful kinds are projectible. In my view, there may be many reasons why we do certain projections and integrate them into our scientific theories.

To summarize, I have claimed that the important question for a theory of scientific kinds is not one about reality, but about how a particular scientific discipline makes projections, i.e. how they pick the properties they hold relevant to their patterns of explanation and classify things accordingly. As such, we can remain agnostic as to whether these taxonomies capture something 'Real', and constrain ourselves to the

study of scientific classification as a study of projection. I have not argued that we should reject realism altogether, but rather than we can bypass the issue and still have an informative account of scientific kinds. If there is still resistance though, I would then let the reader add the adjective 'real' to our scientific taxonomies and move on with our discussion.

4.2.2 Giving up 'natural'

Now that we have established that the question at hand regarding scientific kinds is not their whether they map to a further reality but rather their projectibility, one may ask whether there is any benefit in maintaining the focus on 'natural kinds'. In other words, if the question of what makes certain kinds scientifically interesting need not be formulated in terms of what makes them kinds 'natural' or 'Real' but rather projectible, why keep on talking in terms of 'natural kinds' altogether?

Hacking (2007) raises this issue, and suggests giving up the notion of 'natural kind' altogether. As he puts it:

> Some classifications are more natural than others, but *there is no such thing as a natural kind.*
>
> [...] In the language of classes, there is no well-defined or definable class whose member are all and only natural kinds. Likewise there is no fuzzy, vague, or only loosely specified class that is useful for any established philosophical or scientific purpose, and which is worth calling the class of natural kinds.
>
> Nelson Goodman was right. If the word 'kind' is to be used as a free-standing noun with a grammar analogous to 'set' [...] there are only relevant kinds. (Hacking, 2007, p. 203)

In a similar vein, Ludwig (2018) claims that given the variety of accounts of natural kinds, with their different emphases and focus, there is no interesting use for the notion of natural kind. In his view, we should replace this notion with a multidimensional framework that integrates different proposals and that helps us understand different ways in which science classifies objects.

In contrast, Khalidi (2013) argues that we should retain the term 'natural kind' to separate kinds that capture something in the world rather than arbitrary classifications that we could call 'nonnatural kinds'. Likewise, Ereshefsky and Reydon (2015) claim that we can still talk about natural kinds as those kinds that allow a scientific program to progress and that are empirically testable. In this sense, 'natural kinds' are those kinds with, as they put it, "better epistemic credentials [than others]" (Ereshefsky & Reydon, 2015, p. 984).

I agree, at least in part, with Hacking and Ludwig. In my view, we can drop the adjective 'natural' and instead talk about 'scientific kinds.' As I have argued before, claiming that a particular classification is more real than another does not show why scientists hold it as a preferable basis for projection. The same applies to

the adjective 'natural.' What matters is not whether a kind is natural in contrast to nonnatural, in spite of Khalidi, but rather whether and how it fits into our scientific classificatory schemes. Furthermore, even though I agree with Ereshefsky and Reydon that we should be concerned with how classification aids the progress of a scientific program, I suggest that we still drop the adjective 'natural' to make clear that what we are discussing is scientific classification and not how nature is presumably cut at its joints. The term 'natural kind' does not add much to our current discussion, and hence it becomes a red herring. Consequently, given that the adjective 'natural' does not inform our account of what makes certain classes better suited for projection in scientific inquiry, I will now discuss the issue in terms of scientific kinds, and abandon talk of 'natural' kinds hereafter.

One important consequence of this change in vocabulary is that it makes clear why we ought to hold a pluralistic mindset when discussing scientific kinds. Given that our interest is in how different scientific disciplines classify their objects of study, we need not expect all types of kinds to be successful in projections for the same reason. The motivation behind essentialism and the HPC account lies on what makes certain kinds projectible, but we can still make some room for other types of kinds that are projectible for reasons that fall outside the scope of these accounts. Such could be the case of functional kinds, for instance. What we need in order to understand scientific kinds is not a general theory of what makes all kinds useful, but it might be the case that we need a pluralistic account of what makes different types of kinds meaningful.

In this sense, what we need is a taxonomy of the different types of kinds we use in science, and to elaborate theories of what makes each type useful. Even if we can generalize to some extent and capture a myriad of kinds under a common theory, the end result might involve a set of different theories that account for different kinds. In the last section, I will explore this view, and propose a plausible taxonomy of scientific kinds and examine some prospects for theories that account for each type. This taxonomy may not be exhaustive, but can give us some clues regarding what makes some kinds useful. Moreover, it can give us some idea concerning the kinds that were left out by the accounts mentioned above, and specifically, what sort of kinds could emotions be.

4.3 A taxonomy of scientific kinds

Now we are after a taxonomy of different types of kinds, and exploring what distinctions can be made. The question is: in how many different ways do we justify projection (and therefore induction)? Depending on how many ways we can find, there will be different ways of constructing kinds. In what follows, I will present four plausible types of kinds, and elaborate briefly on what makes them different form the rest. Even if it is the case that there are other kinds available, or that some of the kinds below can be reduced to another type, these will remain open questions for a further theory of scientific kinds. Moreover, it is an empirical matter whether there

are more or less types of kinds available in science, and whether these categories are non-empty or not.

4.3.1 Essential kinds

The first type of kinds are the ones that gave rise to the intuition that natural kinds are individuated by a common essence. These include canonical examples in the literature on natural kinds, such as water and gold. What makes these kinds work for projection is that if we know that they share a common essence, and the possession of this essential property leads to the presence of other properties, then we are justified in projecting the properties we find in a finite observed set of instances to the unobserved set. Consider again the case of water. A projection on water can be phrased as: 'If water has the essence of being a molecule of H_2O, then all instances of water are necessarily molecules of H_2O; this observed sample of water boils at 100°C at sea level in virtue of being a molecule of H_2O; thus, any other unobserved samples of water will be molecules of H_2O and will boil at 100°C at sea level.'

It is worth noting that some objections directed against essentialism imply that essences must be cashed out at different levels, in spite of the tradition's emphasis on physical essences. One such objection comes from the claim that there is no clear criterion to determine how coarse or fine grained our distinctions between essential kinds should be. de Sousa (1984) defends this view. To illustrate this objection, consider the case of water. Water is said to be an essential kind because it is individuated by the essential property of being composed of molecules of H_2O. But as de Sousa points out, H_2O is a chemical and not a physical characterization of water. At the level of physics, there are three stable isotopes of oxygen, ^{16}O, ^{17}O, and ^{18}O.[39] All of these isotopes may be present in water. Thus, at the level of physics, water is individuated by a disjunctive property $H_2^{16}O \vee H_2^{17}0 \vee H_2^{18}O$. Unless we are prepared to accept disjunctive properties as essences, we would be forced to conclude that water is not an essential kind at the level of physics.[40]

What de Sousa's argument shows is that it is unclear at which level should essential properties be cashed out. Hence, the argument invites a multi-level picture of essences. The properties that define essential kinds and that we may therefore be willing to call 'essential' depend on which level the generalizations are being made. In this view, water is an essential kind defined by the property of being H_2O even if this property has different possible realizers at lower levels. What makes water an essential kind is that the projections that are made across members of the kind are made in virtue of them possessing the essential property. To put things differently, since the question of how fine- or coarse-grained our distinctions should be depends on what we want to explain, the characterization of the essence of an essential kind also depends on

[39] de Sousa only mentions two isotopes, ^{16}O and ^{18}O, which are the most common ones. Chemical peculiarities aside, his argument still stands.

[40] A similar argument is proposed by Churchland (1985). Churchland argues that if we restrict ourselves to kinds that figure in basic laws, then kinds such as water would not be natural (or in my terms, essential). This is because water as such does not figure in any basic law, but only atoms or more fine-grained micro-structures.

how is the kind explanatorily relevant for a discipline. Water characterized as H_2O is relevant for chemistry, although I might not be for micro-physics.

The question remains how to characterize essences as such. I will not solve this issue here, as this would imply dwelling into deep metaphysics (for a discussion see Roca-Royes (2011)). However we characterize them, what is important for our purposes is that essential properties may provide a basis for projection in cases where they explain the presence of other interesting properties members of the kind possess. As presented above, water's being H_2O explains why it boils at 100°C at sea level, why it freezes at 0°C, and other properties that we observe. Hence, what we need is an account of essences as special properties that explain other properties of the kind.

4.3.2 Historical kinds

The second type of kind are kinds individuated by their common history. These are the sort of kinds that defenders of the HPC account often appeal to, the paradigmatic cases being species and biological taxa. What individuates them, as seen in the case of lilies, garlics, and onions, is that members of the kind have a common causal history. In the case of species, this may be cashed out as a common evolutionary history. In other cases, a common causal history may be cashed out as common design (in the case of artifacts such as chairs, for instance) or common social development (as Bach (2012) claims is the case of gender).

Historical kinds contrast with essential kinds, not only in that they are projectible in virtue of their history rather than some essence, but also because, as Millikan pointed out (although using the language of 'natural kinds'), what individuates them is dependent on the actual course of things, rather than an eternal property independent of spatiotemporal considerations. In the case of biological species, what makes an organism a member of species is sharing the same evolutionary history with other members, and this history depends on the course of states of affairs that has been the case in our actual world. This involves considering how things actually developed rather than adopting an ahistorical perspective. The traits that are explanatorily interesting in the case of species are thus dependent on how things unfolded in our particular causal history.

To understand how projections based on historical essences might work, consider the case of another biological species: tigers. We can expect other tigers to have stripes because they have a shared evolutionary history. Moreover, even if a particular tiger is born without stripes, it would still count as a tiger because it still shares the same history and may share other traits such as sharp teeth and claws because of this evolutionary history. In this sense, we project from one member of the kind to others because sharing a common causal history can be taken as a good reason to expect other unobserved members of the kind to have certain properties.

4.3.3 Functional kinds

The third type of kinds in our taxonomy are functional kinds. These are the kinds that, as I argued above, were left out by essentialist and HPC accounts, and they include kinds in psychology, neuroscience, but also some biological (e.g. predator), chemical (e.g. water[41], and physical kinds (e.g. machine). They are characterized by allowing multiple realization in different systems, and they owe there unity to the functions they carry out.

Functional kinds are among the most problematic type of kinds, some even denying that they may form interesting scientific kinds at all. The reason behind these doubts is that properties of kinds that allow multiple realization cannot be projected, because even if two systems realize the same function, they may differ in important properties that may preclude one system to have the same properties as the other. This was the sort of objection put forward by Millikan and that lies at the heart of resisting the view that functional kinds are natural in any sense. In my view, we can resist these objections and propose a scientifically interesting account of functional kinds by recognizing that functional kinds are not projectible in virtue of any arbitrary property, but rather in virtue of a property at a higher level of description. These higher order properties provide interesting explanations and are not merely placeholders waiting to be replaced by lower level properties. Hence, functional kinds are also scientific kinds in their own right.

In the discussion above, I suggested that functional kinds should be considered scientific kinds in their own right in virtue of their functional unity. Kinds such as psychiatric categories, neuroscientific phenomena, etc., have proven to be scientifically useful for the disciplines in which they figure. Consequently, instead of maintaining one theory for all kinds that excludes functional kinds, I claim that we should formulate a theory that accounts for the usefulness of these kinds, even if it turns out to be a different theories than for the other types.

One classic way of framing the problem is in terms of reduction laws, as we find in the discussion between Fodor and Kim. As explained above, Fodor (1974) claims that kinds in the special sciences (i.e. functional kinds) are not reducible to lower level kinds, and therefore they constitute autonomous kinds. The reason is that since functional kinds are multiply realizable, higher level laws connecting functional kinds of the form $M \rightarrow Q$ end up reducing to disjunctions such as $(P \vee Q \vee S) \rightarrow (T \vee U \vee V)$. Fodor thinks that the higher level laws are scientifically interesting in their own right, and hence he takes this to be an argument for the autonomy of the special sciences.

Kim (1992), just like Millikan, is skeptical about this argument. In his view, if Fodor is right that functional kinds do not reduce to lower level kinds, then we ought to eliminate them rather than keep them as autonomous kinds. This is because the functional kinds involved in psychology and other special sciences, which are the ones Fodor has in mind, reduce not only to disjunctions but to predicates that are *wildly*

[41] Water can be said to be a functional kind if we buy the idea discussed above that kinds defined disjunctively at the physical level are multiply realized in different physical entities, such as different isotopes of H_2O.

disjunctive or *heterogeneous*. These predicates, Kim thinks, are not projectible, and therefore cannot be taken as scientifically interesting predicates. As he puts it:

> There is nothing wrong with disjunctive predicates as such; the trouble arises when the kinds denoted by the disjoined predicates are heterogeneous, "wildly disjunctive", so that instances falling under them do not show the kind of "similarity", or unity, that we expect from instances falling under a single kind. (Kim, 1992, p. 13)

The discussion between Fodor and Kim sheds light on the nature of functional kinds. Kim and Fodor both recognize that functional kinds appear to be irreducible to lower level kinds (e.g. physical kinds), but draw different conclusions. For Fodor, these kinds are projectible and well-entrenched, whereas for Kim they are not. How should we decide?

Kim and Fodor's discussion is framed in the context of an axiomatic view of theories along with a nomological model of scientific explanation. In this context, scientific theories are characterized as a set of laws spelled out in a basic vocabulary and an interpretation connecting terms in the basic vocabulary to entities in the world. Multiple realization becomes problematic because higher level laws don't reduce to lower level laws without recourse to disjunctions. Since theories to be reduced to a basic vocabulary, the irreducibility of functional kinds turns out to be an issue.

Given that the problem Fodor and Kim are discussing depends on our acceptance of this picture of science, one possible move is to reject the picture altogether. This is what Klein (2013) argues for. Klein holds that the traditional problems with multiple realizability stem from the axiomatic view of theories, and once we abandon the axiomatic view, the problems dissolve. He suggests adopting a semantic view of theories instead, a view in which theories are taken as a family of models with which we represent the world. By doing so, we can account for functional kinds as being kinds that we pick out in models in which realizers are irrelevant for the representational purpose of the model. In the case of psychological kinds, for instance, they are just the kinds that figure in models that represent an our mental lives. Given that this picture does not demand reduction to a basic vocabulary, the problems that Fodor and Kim discussed do not arise, at least in principle.

Shifting towards a modeling view does relieve some pressure, but this still does not save functional kinds from being troublesome. Piccinini and Craver (2011) claim that analyzing cognitive capacities in terms of functions does not lead to a full-fledged explanation of the phenomenon, but rather to a *mechanistic sketch*. As such, classification in terms of functional kinds is merely an approximation towards explaining phenomena, but actual explanation comes only when we complete the explanation with an account of the mechanisms realizing these functions. In other words, functional kinds are not genuinely explanatorily interesting kinds, but only placeholders in an incomplete explanation.

In my view, the question is not whether functional kinds are scientifically useful but rather how could they be so. As I argued above, there are a number of kinds in

psychology and psychiatry that are individuated functionally and whose characterization may disregard details about their realizers. Moreover, I also mentioned kinds in other sciences that are individuated functionally, such as «predator» in biology and «machine» in physics. What we need is then an account of what makes functional kinds projectible if its not a property of the realizers or their causal history.

In general, we can understand functional kinds as kinds united by a given function or capacity, i.e. a kind whose members contribute in similar ways to a system. Hence, to identify functional kinds, we must first engage in *functional analysis*. In what follows, I will adopt the account of functional analysis put forward by Roth and Cummins (2017). This account has the virtue of allowing an account of functional kinds while tackling some of the issues above and being descriptively adequate, at least for the purposes ahead. Given that a thorough discussion of functions and their role in science requires a much deeper treatment, I will assume Roth and Cummins's general account without further discussion. A similar account can be found in Weiskopf (2011).

Roth and Cummins think of functional analysis as follows:

> Functional analysis is the attempt to explain the properties of complex systems [...] by the analysis of a systemic property into organized interaction among other simple systemic properties or properties of component subsystems. This explanation-by-analysis is *functional* analysis because it identifies analyzing properties in terms of what they do or contribute, rather than in terms of their intrinsic constitutions. (Roth & Cummins, 2017, p. 35)

In terms of kinds, functional kinds are kinds formed by similarities in terms of what a group of objects do or contribute, i.e. what functions they have. They disregard details about their realizers insofar as these details are irrelevant for the explanation of the capacity these objects have. As Roth and Cummins formulate it, functional terms (which would pick out functional kinds) act as *causal relevance filters* that abstract away from particular properties of the realizers and instead highlight a system's functional features. This makes functional kinds multiply realizable and hence different from essential and historical kinds.

It is worth clarifying that this account need not imply that realizers are irrelevant in general. Instead, the claim is that asking what the realizers of a given function are implies asking a completely different question than those concerning how a system carries out said function. It is quite plausible that we can reach a more thorough understanding of certain systems by looking at the realizers (i.e. brains as realizing psychological states), but this does not imply that functional analyses are incomplete or that we need details about the realizers in every case. Moreover, as Sullivan (2016) points out, the functional characterization of a capacity is often prior to the search for its neural basis or its cellular and molecular mechanisms (in the case of neuroscience). In this sense, functional analyses and functional kinds are genuinely interesting scientific strategies and classifications in their own right.

This view of functional kinds as kinds formed by functional analysis helps us tackle the problems above. In this view, functional kinds are not placeholders in an incomplete explanation, but members of explanations that differ from the sort of mechanistic explanation Piccinini and Craver have in mind, one in which details about the realizers are vital.[42] Furthermore, there is no pressure to have them figure in laws reducible to lower levels because, first, the model of explanation at play is mechanistic rather than nomological, and second, because functional explanations are taken as explanatorily interesting in their own right without recourse to lower level realizers.

4.3.4 Social kinds

Lastly, we have the so-called social or human kinds. These include kinds such as «money», «marriage», «racism», and the like. On a first approximation, we can think of social kinds as kinds united by certain forms of interaction and the presence of an institutional framework that gives meaning to these objects. Instances of money are material (or even virtual) objects which we exchange for goods and services, marriage is an activity where individuals declare a particular type of relationship, racism is a systematic oppressive attitude towards a group of people in virtue of their race, etc. What all of these have in common is that what unites members of the kind is the sort of social world in which they acquire their meaning and role.

At a first glance, social kinds appear to be very different from at least essential and historical kinds.[43] What unites social kinds is not a property external to ourselves, and hence they are in a sense not independent from us as human beings. This gives off the impression then that social kinds are not natural kinds at all, but rather arbitrary and mind-dependent categories. As such, they seem to differ in important respects from the sort of kinds the tradition had in mind when developing accounts of natural kinds.

In spite of this skepticism about social kinds, it is false that classifications in these terms are arbitrary. The social sciences use social kinds to group phenomena which they find relevant and explanatorily interesting. For example, while it is true that, strictly speaking, any physical object can be used in a transaction and hence be used as currency, it is not the case that the category CURRENCY is arbitrary. Objects used as currency are objects that are used in a specific set of social practices and whose dynamics are worth studying. In this regard, social kinds such as currency allow generalizations and projections such as the ones we find in the social sciences.

One way in which we can start characterizing social kinds is following Searle's (1995) account of social facts. Searle starts by distinguishing two senses in which we

[42] It is still unclear whether functional explanations are categorically different from mechanistic explanations. If mechanistic explanations are explanations in terms of entities and activities, then it could be argued that the entities in a mechanistic explanations may as well be abstract, functionally characterized entities.

[43] It is plausible that we can subsume social kinds under functional kinds if we adopt a broad account of functions. Insofar as social interactions and institutions could be described functionally, social kinds may be a subtype of functional kinds. At the very least, this is not obviously false. I will, however, treat them as different types of kinds in what follows.

can talk about objectivity and subjectivity, namely, an epistemic sense (which applies to judgments) and an ontological sense (which applies to entities). In the epistemic sense, a judgment is *epistemically subjective* when its truth or falsity depends on the attitudes, feelings, and points of view of the makers and hearers of said judgment, and it is *epistemically objective* in the opposite case. For example, my judgment that 'Coffee is delicious' is epistemically subjective insofar as it's truth depends on my own attitudes towards coffee. Likewise, my judgment that 'Coffee is an acidic drink' is epistemically objective, as it is true as a matter of fact and not because of our own attitudes. As for the ontological sense, an entity may be *ontologically subjective* in case its existence depends on the attitudes, feelings, and points of view of a perceiver, and *ontologically objective* in case it does not. To use Searle's example, pains are, in this view, ontologically subjective, as they require a perceiver to exist, whereas mountains are ontologically objective insofar as they would exist regardless of our own existence.

What Searle points out with this distinction is that social facts are ontologically subjective but epistemically objective. They are ontologically subjective in the sense that they require the existence of subjects to exist in the first place, but our judgments about then are true not in virtue of our own attitudes but rather they are true in virtue of the states of affairs of our social world. A fact such as «Transgender women in immigration detention in the US are often victims of sexual assault and denial to necessary medical care» (Frankel, 2016) is ontologically subjective insofar as the existence of transgender women, immigration, detention, etc., requires the existence of subjects in a society with an institutional framework that gives meaning to the notions of sexual and gender identity, political borders, etc. However, it is epistemically objective as its truth does not depend merely on us thinking this is the case or not. Rather, it becomes a fact of our social world, and as such its truth conditions depend on a particular social state of affairs.

If we accept Searle's distinction, it would follow that social kinds are kinds united in virtue of ontologically subjective, epistemically objective facts about our social world. The kind «transgender women», for example, is comprised of the group of people whose gender identity was assigned as male at birth and who have transitioned into a different gender role, namely, that associated with womanhood. As a kind, it allows us to do some generalizations and projections that enable us to understand a range of social facts. Under Searle's taxonomy, it is an ontologically subjective kind in the sense that without human beings along with our historical attitudes towards gender, the kind would not exist. Yet, our judgments about the kind are epistemically objective in that they do not depend on what we think about the kind, but on social facts about it.

Searle's account provides an interesting starting point towards understanding social kinds, but it does not cut it as an overall account. As Khalidi (2013) points out, following Thomasson (2003), Searle's account is too restrictive and only applies to a subset of social kinds. While presenting his account, Searle has in mind kinds that are indeed dependent on people having thoughts about them for their existence, such as money or marriage. These do not exhaust social kinds though. Khalidi considers

Thomasson's example of «recession» as well as «inflation», «racism», and «poverty». All of these constitute social kinds insofar as they depend on a social and institutional framework, but they do not depend on the existence of thoughts about them. Even before the development of economics as a discipline there was poverty and inflation, and even before we constructed a concept to capture racism there was discrimination in virtue of race.

Khalidi proposes a three-category taxonomy of social kinds, building on the argument above. The first type of social kinds are mind-dependent kinds in which human mental attitudes need to be in place for the kind to exist but they need not be directed towards the kind itself. This is the case of «recession» and «racism», kinds for which certain attitudes towards commodities or racial groups need to be in place for the kind to obtain, but that do not require human beings to be conscious of the existence of the kind.[44] The second type are mind-dependent kinds for which mental attitudes need to be in place but that can be instantiated without attitudes towards a particular instance. Khalidi illustrates this type with the case of «war». In order for there to be wars, there need to be explicit attitudes towards wars, but an instance of a war can obtain even if we don't think about it as a war (perhaps we think of it as a mere criminality problem or a set of battles rather than a war). Lastly, there are mind-dependent kinds whose existence and instantiation both depend on mental attitudes. This is the case of «permanent resident», which requires someone believing a person has resident status in order to exist and be instantiated.

I find Khalidi's taxonomy of social kinds plausible, although in his view, the third type of social kinds do not constitute natural (in my case scientific) kinds. He claims so because this type of kinds do not reflect causal patterns. The properties associated with these kinds are not associated with it as a matter of causality but are instead codified into the kind as a matter of explicit convention or law. Consider again the kind «permanent resident». Properties associated with this kind depend on how a given country defines the kind and the conditions whereby a person becomes a member of it. Since Khalidi thinks that the presence of causal patterns is necessary for natural kindhood, he excludes these kinds from the list of scientifically interesting kinds.

In my account, however, we need not exclude them. They are kinds whose projectibility depends on explicitly coded factors, but this need not imply that they are not scientifically interesting. Politologists, sociologists, anthropologists, and many other social scientists, to mention a few, may find a number of interesting facts about these kinds and generalize successfully. It is also plausible that other scientific disciplines find important generalizations based on social kinds, as can be seen in research in social psychology, evolutionary biology, and the like. If this is true, I submit that social kinds do make it into the list of interesting scientific kinds.

[44] Guala (2014) criticizes Khalidi for claiming that all social kinds are mind-dependent. He argues that the presence of attitudes directed towards the kind itself is neither necessary nor sufficient for social kinds to exist. As Khalidi presents his account though, the criticism seems misguided, as Khalidi clearly accepts the existence of social kinds that do not require such attitudes towards the kind itself.

To summarize, social kinds are those whose existence and projectibility depend on the social and institutional framework in which we live. Some of them are conventional insofar as they are defined by a set of rules or laws, but others need not be conventional in this respect and can even exist without people having beliefs and attitudes towards the kinds themselves. Furthermore, they are kinds of which we can make judgments that are, as Searle puts it, epistemically objective, in the sense that we hold these judgments true as a matter of a given state of affairs rather than merely subjective attitudes.

4.4 Conclusion

In this chapter, I have presented an account of scientific kinds that offers us different alternatives to answer the question of what type of kind are emotions. I started with the question of why are only some kinds scientifically interesting, and I presented some traditional answers to this question following the tradition of Mill, Kripke, Putnam, and Boyd. I accused this tradition of identifying scientific kinds with only one type of natural kind (essentialist or HPC), hence attempting to pass one account of natural kindhood as an account for all scientific kinds.

Instead of proposing one overall account of natural kindhood to explain scientific kinds, I proposed tracing back what made scientific kinds interesting. I suggested that what constitutes scientific kinds is that we can use them in projections and generalizations. This opened up the possibility that different scientific kinds may be projectible for different reasons, and thus that there may be different types of scientific kinds identified with different reasons for projection. As a result, I argued that we need to look into different projectibility patterns in science, rather than their correspondence with some ultimate taxonomy or Reality. Lastly, I presented four types of scientific kinds: essentialist, historical, functional, and social kinds.

With this taxonomy of scientific kinds at hand, we can now move forward to the question: what type of kind are emotions? To tackle this question, we need to make clear what sort of properties are associated with emotions and what makes the category projectible in the contexts in which it figures. To do this, we must find a way to characterize the phenomenon to be explained and subsequently construct scientific concepts that will allow their integration into a scientific theory. In other words, we must *reconstitute the phenomena*, as Bechtel and Richardson (2000/2010) put it. In the next chapter, I will present reconstitution and argue that in the case of emotions, we can carry out reconstitution by explicating folk emotion concepts. Once we have clarified this strategy, I will apply it in chapter 6 and argue that emotions are best understood as functional kinds.

Chapter 5

Taking a step back

Reconstitution through explication

There are times in the course of the history of science when researchers must reconsider their conceptualization of the phenomena to be explained. Maybe they have not identified different kinds of phenomena, misclassifying them under a common category. Maybe they have thought that a single phenomenon was present when there were actually two different things at place. Or maybe they have not found a level of analysis that would allow them to see the behavior they are interested in investigating. In any of these cases, scientists need to take a step back and rethink the phenomena they are after. In the literature on mechanistic explanation, this is what Bechtel and Richardson (2000/2010) call *reconstituting the phenomena*.

In the previous chapter, I concluded that in order to overcome the challenges that emotion research faces, we need to decide which type of kinds do emotions best conform to. To do this, we need to make clear the type of inductive inferences we are interested in making when it comes to emotions. Given that there are several options available to us, we need some criteria to make this decision.

In this chapter, I will discuss reconstitution as a strategy to carry out such a task. I will argue that given the past difficulties to identify emotion kinds, we need to take a step back and reconceptualize the explanandum phenomenon at play in emotion research. To do this, I start by introducing Bechtel and Richardson's notion of reconstitution in the framework of mechanistic explanation. I examine their case study in genetics and discuss some of its consequences for a general philosophy of science. Yet, this will only constitute the first part of my argument in this chapter.

In the second part of the chapter, I will consider some peculiarities when it comes to concepts such as those about emotions. In my view, emotions, as well as other related phenomena in psychology, are first encountered as part of our everyday practices. In other words, emotions constitute folk phenomena, and we make reference to them by the use of folk concepts. So far, this is not surprising. What is interesting, however, is that this has consequences when it comes to reconstitution. When the phenomena that we need to reconstitute are folk phenomena, then we have a clear strategy to carry out reconstitution, namely, by analyzing folk concepts as ostensive devices to pick out the

phenomena and working on their basis to construct scientifically adequate concepts. In this sense, reconstitution becomes a matter of *explicating* folk concepts. Put slightly differently, in the case of folk concepts, reconstitution takes the form of *explication*.

If my view is correct, this has direct consequences regarding the question of what type of kinds emotions constitute: To make a decision about the type of scientific kinds emotions are is to find the best level of analysis on which we can reconstitute and explicate emotion concepts. By seeing reconstitution and explication as strategies to pick out the correct scientific kinds, we end up at the gates of my final claim, i.e., that the best way to reconstitute and explicate emotion concepts is by invoking functional kinds.

As I already explained, I will start off by introducing Bechtel and Richardson's account of reconstitution. After discussing their case study of Mendelian genetics, I draw some preliminary conclusions about the nature of reconstitution. Afterwards, I discuss explication as a form of reconstitution. I introduce the traditional Carnapian account of explication and discuss some recent modifications made to the account. I conclude by examining the consequences of these ideas when it comes to the problem of identifying emotion kinds.

5.1 What is reconstitution?

5.1.1 Reconstitution and mechanistic explanations

Reconstitution, on a broad construal, consists of the *reconceptualization* of a phenomenon to be explained. Bechtel and Richardson (2000/2010) introduce reconstitution in the framework of mechanistic explanation. Bechtel and Richardson are interested in how scientists conceptualize phenomena in a way that enables localizing and decomposing them into components and interactions that figure in mechanistic explanations. Yet, much of what they say about reconstitution can be applied independently of the mechanistic framework. What I am interested in is in their remarks about when and how scientists reconceptualize phenomena in order to advance a scientific research program. Still, let us introduce reconstitution in terms of mechanisms and later generalize to general attempts to reconceptualize the phenomena.

Generally speaking, and at the risk of sounding circular, a mechanistic explanation is an explanation of a phenomenon in terms of mechanisms. A mechanism, in turn, can be characterized in different ways. As Craver and Tabery (2017) explain, there are three often cited characterizations of mechanisms:

MDC "Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer, Darden, & Craver, 2000, p. 3).

Glennan "A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interaction between

parts can be characterized by direct, invariant, change-relating generalizations" (Glennan, 2002, p. S344).

Bechtel and Abrahamsen "A mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena" (Bechtel & Abrahamsen, 2005, p. 423) (adapted from Craver & Tabery, 2017)

In all three characterizations, a mechanism is defined as a collection of entities or parts arranged in a particular way such that they interact in order to function or produce some phenomenon. A simple and familiar example of a mechanism presented by Glennan (1996, pp. 56-58) is a float valve found in a regular toilet. A float valve is a mechanism that regulates the water level inside a tank (phenomenon). It consists of a float and a lever that opens or closes an intake valve. When the water level is down, the float drops and causes the lever to go down. This in turn opens the intake valve, allowing water to fill the tank. When the water level rises, it raises the float and causes the lever to go up, closing the valve and stopping further water intake.

According to defenders of the mechanistic account, explaining how the float valve amounts to indicating how the entities (float, lever, valve, and water) interact (causing the float to lower or raise, causing the lever to go up or down, etc.) in order to produce the phenomenon in question, i.e., the regulation of the water level. Hence, mechanistic explanations depend on how the phenomenon is identified. Since mechanisms are always mechanisms producing a phenomenon, characterizing the phenomenon adequately is a necessary condition for a successful mechanistic explanation. I will come back to this below.

Mechanisms and mechanistic explanations are ubiquitous in science, as New Mechanists stress. Consider Craver's (2007) example of the explanation of neurotransmitter release:

> The mechanism begins, we can say, when an action potential depolarizes the axon terminal and so opens voltage-sensitive calcium (Ca2+) channels in the neuronal membrane. Intracellular Ca2+ concentrations rise, causing more Ca2+ to bind to Ca2+ / Calmodulin dependent kinase. The latter phosphorylates synapsin, which frees the transmitter-containing vesicle from the cytoskeleton. At this point, Rab3A and Rab3C target the freed vesicle to release sites in the membrane. Then v-SNARES (such as VAMP), which are incorporated into the vesicle membrane, bind to t-SNARES (such as syntaxin and SNAP-25), which are incorporated into the axon terminal membrane, thereby bringing the vesicle and the membrane next to one another. Finally, local influx of Ca2+ at the active zone in the terminal leads this SNARE complex, either acting alone or in concert with other proteins, to open a fusion pore that spans the membrane to the synaptic cleft. (Craver, 2007, pp. 4-5)

Details and addenda aside, the explanation presented by Craver is a clear example of a mechanistic explanation: it invokes entities and their interactions such that they produce a given phenomenon. More specifically, the explanation presented here is an explanation of the phenomenon of neurotransmitter release in terms of different parts of the neuron, chemical compounds, and their interactions in terms of activities of the membranes and chemical reactions.

In order to identify mechanisms, scientists employ a number of heuristic strategies. Bechtel and Richardson (2000/2010) introduce two such strategies, namely, the aforementioned decomposition and localization strategies. Decomposition consists of subdividing the explanatory task in separate parts to make it more intelligible. In the first example above, we may decompose the action of the float valve into opening and closing the intake valve, the raising or lowering of the lever, and the floating of the float. In the second, we may divide the overall synaptic mechanism into the binding of Ca2+ to Ca2+/Calmodulin dependent kinase, the incorporation of v-SNARES into the axon terminal membrane, and the like. Each of these actions can be studied as a mechanism in their own right, i.e., as consisting of entities and interactions that explain how a given phenomenon obtains. Hence, we can localize it as carried out by a subset of the entities involved in the overall mechanism (e.g., raising or lowering the level as localized in the float valve or binging of Ca2+ to Ca2+/Calmodulin dependent kinase in the synaptic receptor).

When decomposition and localization fail, Bechtel and Richardson argue, scientists pursue reconstitution. This happens when scientists believe that they have had identified the phenomenon to be explained wrongly, precluding us from adequately decomposing the phenomenon into components and their interactions and localizing them appropriately. As a result, when failure ensues, we need to reconceptualize the phenomenon produced by the mechanism, i.e., we need to *reconstitute* the phenomena.

5.1.2 Bechtel and Richardson on Mendel

Bechtel and Richardson present reconstitution through a discussion of Mendel's (1865/2009) findings on inheritance. Prior to Mendel, ideas about inheritance were present, but a proper theory of inheritance was lacking. Mayr (1982, p. 634) mentions some of these prior ideas, including that only one of the parents transmits elements to the offspring, that the contributions of the father and the mother are quantitatively and qualitatively different, or that both contributions are blended into an average. With Mendel's studies, not only did genetics receive an initial formulation of currently accepted laws of inheritance, but also a paradigm that would constitute the basis of future work.

Mendel's experimental paradigm is rather simple. He crossed a number of different varieties of peas with each other, keeping track of how different traits were passed on across generations. Initially, Mendel observed seven different traits, including the form and color of the seeds, the position of the flowers, and the length of the stem, among others. He noticed that among different pairs of traits (e.g. yellow vs. green seeds), the first generation exhibited the traits in a 3:1 proportion. In other words, for every three plants with yellow seeds, there would be one with green seeds. This led him

to claim that one trait would have priority over another one, hence coining the terms 'dominant' and 'recessive,' respectively.

In the second generation, however, things were different. Among the plants that were bred from plants exhibiting dominant traits, there were now offspring that exhibited recessive traits. Specifically, Mendel saw that among the offspring of dominant-exhibiting plants, one third exhibited these recessive properties. Consequently, the overall proportion in the second generation as 1:2:1 (for every four plants, one had pure dominant traits, two had mixed traits, and one had pure recessive traits). Further generations would continue to obtain in this proportion, suggesting a common pattern.

Mendel's findings yielded what later came to be known as Mendel's laws of inheritance. Bechtel and Richardson focus on the first two laws, the law of segregation and independent assortment:

Law of segregation Each parent passes on only one allele [a variant of a given gene] to each offspring.

Law of independent assortment How alleles segregate at one locus is independent of how they segregate at another locus (adapted from Griffiths & Stotz, 2013, p. 14).

Of special interest to Bechtel and Richardson's discussion is the law of independent assortment. To understand its importance, it is vital to go over some clarificatory notes. The notion of «allele» and—more importantly—«gene» were not part of Mendel's vocabulary (nor was the term «law» for his conclusions). In this sense, we must first approach «gene» as a hypothetical or a black-boxed unit. A gene in this sense is merely whatever transmits a trait from the parent to the offspring.[45] Similarly, an allele must be understood as whatever determines a given trait, and presumably a part of a gene. On this initial interpretation, the laws state that whatever passes on to the offspring is only one determinant of a trait (law of segregation) and this passing on is independent from other passing-on's that might occur (law of independent assortment). This construal will help us understand the problem with these laws later on, and the subsequent reconstitution as construed by Bechtel and Richardson.

As Bechtel and Richardson explain, these laws led to the assumption, or were based on the assumption, that one gene would correspond to one trait. This is commonly referred to as the 'one gene-one trait' hypothesis. What is crucial to Bechtel and Richardson's discussion, and what introduces reconstitution, is that in the years following the rediscovery of Mendel's work, findings did not match the expectations yielded by this assumption. This led to reconceptualizing the phenomenon at stake when it comes to formulating a theory of inheritance, i.e., to reconstitution.

[45] Mendel talked about each reproductive cell being "provided with the material for creating quite similar individuals." See Mendel (1865/2009, p. 67)

The *Drosophila* challenge

After presenting the background of Mendelian genetics, Bechtel and Richardson introduce one of the first important challenges: Thomas Morgan's studies with *Drosophila megalonaster*, a species of flies. In the early twentieth-century, Morgan conducted a number of experiments crossing *D. megalonaster*. He noticed a male mutant variant that exhibited white eyes, and crossed it with a wild-type red-eyed female. In the first generation, all of the offspring exhibited red-eyes, while in the second generation, Morgan observed the 3 red-eyed to 1 white-eyed ratio Mendel had observed in the peas before. Interestingly, when including whether the offspring was male or female, Morgan noticed that all white-eyed offspring were male, even though the distribution between males and females was balanced. This suggested that there might be a relationship between sex and the mutant-variant gene, a result that contradicted independent assortment. If independent assortment was true, sex and white-eyes would have to be independent from each other.

A student of Morgan, Alfred Sturtevant, later found more evidence against independent assortment as previously understood. Bechtel and Richardson discuss two of his findings, both coming from experiments with *D. megalonaster*. First, Sturtevant showed that the eye color of the *D. megalonaster* could be altered under the influence of surrounding tissue. If independent assortment were true, the expression of a given eye color would have to be independent from these kind of factors. Second, the researcher found that in some variants of the *D. megalonaster*, genes were brought into adjacent positions inside the chromosome (by this time, researchers had already localized genes in the chromosome; see Darden & Maull, 1977). When this happened, individuals would exhibit different characteristics than others with the same genes in different positions. Again, if independent assortment were true, the position of a gene inside the chromosome would have no bearing on its expression.

These findings, as discussed by Bechtel and Richardson, challenged the idea that genes worked independently of each other. Furthermore, they also challenged the 'one gene-one trait' hypothesis. The presence of a given gene would no longer predict the presence of a trait, since the presence of a trait was now found to depend on other factors (sex, surrounding tissue or the position of the gene in the chromosome). In this sense, the theory of inheritance proposed by Mendel and developed by his followers did not explain traits. Put differently, decomposition of inheritance into genetic factors directly altering traits, and the subsequent localization of these relations, fell under pressure. This forced researchers to find a different level of analysis, a different level at which genes acted and that would be the phenomenon explained by the Mendelian theory of inheritance. As Bechtel and Richardson show, this was the level of enzymes.

From genes to enzymes

After the studies of Morgan and Sturtevant, scientists continued studying inheritance in more detail. One particularly important study is the one by George Beadle and Edward Tatum. Beadle and Tatum studied *Neurospora crassa*, a type of bread mold.

Given the paradigm Mendel had developed and that had become the standard in genetics, their study is straightforward. The basic idea was to obtain different mutant variations of *N. crassa* and observe how different variations in their genes led to different traits. In this case, however, they induced mutations using X-rays (rather than waiting for a mutation as in the case of *D. megalonaster*).

After obtaining a number of mutant variants, the researchers transferred the specimens into minimal environments on which wild-type variants could thrive. Due to the mutation, a number of mutant variants were not able to survive in this minimal environment. According to the researchers, this was because the mutants were not able to synthesize some substance necessary for their survival, a substance wild-type variants could produce. However, once the researchers added different substances to the minimal medium, the mutant *N. crassa* were able to grow. Some of these mutant variants survived in a medium supplemented by arginine; some by either arginine or citrulline; and others by arginine, citrulline, or ornithine.

At a first glance, all that was needed for *N. crassa* to survive was a simple two-step process: wild-type variants synthesized ornithine into citrulline, and citrulline into arginine, which warranted survival. Hence, in the mutant variants, one of these reactions was impeded. If the first reaction was blocked, adding either citrulline or arginine to the medium would suffice for the mutants to survive. If the second reaction was blocked, arginine was required. Finally, if the synthesis of ornithine was blocked, either of the three substances was sufficient.

Even though the process was more complex than presented above, since arginine could also be turned into ornithine and urea, the moral of the story is the following. By intervening on specific genes, Beadle and Tatum showed that what was blocked was the synthesis of a specific enzyme, which in turn could or could not affect the synthesis of other enzymes and subsequent proteins, ultimately affecting some trait (in this case, surviving). As a result of these experiments, the 'one gene-one trait' hypothesis shifted to a lower level of analysis, becoming the 'one gene-one enzyme' hypothesis.

Reconstituting the phenomena

The shift from the 'one gene-one enzyme' hypothesis, as presented by Bechtel and Richardson, offers an interesting case study for reconstitution. At the beginning of the story, scientists pursued an assumption regarding the phenomenon to be explained by genetics, namely, the direct influence of genes on expressed traits. Later, through a shift in the level of analysis, scientists reconceptualized this phenomenon, now cashing out the influence of genes at the level of enzymes. This move rescued the law of independent assortment, which now could be taken to hold at the level of enzymes, and allowed further progress in genetics. In Bechtel and Richardson's words:

> Genes are thought of as specific in their action and as acting in relative independence of other genes. There is independence at the level of the enzymes they produce. Conceived in terms of the observable phenotypes, genes have

> a complex role in the development and metabolism, and a complex organization. Simplification followed only on understanding the phenotype differently. A characterization of the phenomena in terms of observable traits was replaced by one couched in terms of biochemical products. This was still localization and decomposition, but this time the reconstitution of the phenomena with a shift to a lower level allowed us to retain localization and decomposition in the face of complex organization. (Bechtel & Richardson, 2000/2010, p. 194)

As the discussion of the development of genetics shows, scientists often do not give up their hypotheses in presence of conflicting evidence. Rather, they revise some of their assumptions, even reconceptualizing what they thought was the phenomenon to be explained in order to preserve other parts of their theories and research programs. In other words, there are times in science where what we need is to rethink what is it that we wanted to explain, and find other ways of describing the phenomenon that allow us to advance our research. This, I take it, is what Bechtel and Richardson have in mind when it comes to the notion of reconstitution.

As informative as Bechtel and Richardson's discussion might be, the overall notion of reconstitution is still unclear. Specifically, we can raise the following questions: how do scientists carry out reconstitution? That is, how do they determine the level of analysis that allows them to reinstate localization and decomposition? Additionally, is reconstitution always linked to mechanistic explanations? In order to answer these questions, let us discuss reconstitution in more detail.

5.2 Examining reconstitution

To approach the question of what are the conditions for a successful reconstitution, let us first observe that reconstitution is a broader strategy than the one presented by Bechtel and Richardson. In the case of genetics, reconstitution consisted in a reconceptualization of the phenomenon by shifting to a lower level of analysis. Nevertheless, as Kronfeldner (2015) argues, reconstitution can also happen by shifting to a higher level, i.e., by abstraction. If this is true, then this raises the issue of determining whether we should move to a lower or a higher level of abstraction in presence of problematic findings.[46] Let us elaborate on this claim.

Kronfeldner makes her case by considering an example of an explanation of human height, and a hypothetical controversy that illustrates how this shift could also qualify as a form of reconstitution. Imagine two scientists discussing data on changes in human height since the fifteenth century. They observe that height has increased throughout the centuries, and that males tend to be taller than females at any given point within

[46] Furthermore, it also suggests that reconstitution might not be intrinsically linked to mechanistic explanations, at least in terms of lower level mechanisms. Whether or not we would call higher level explanations mechanistic is a topic that lies outside the scope of the present discussion, however. Since the mechanistic framework allows different levels of mechanisms, I will assume that moving to a higher level of abstraction still yields a mechanistic explanation. For a discussion on levels, see Craver (2007, ch. 5)

this timeframe. Facing these phenomena, the scientists raise the question of what explains the data at hand.

For the first scientist, a biologist, the data are better explained in terms of genetics. This is because they observe that there are differences between males and females that can be explained in terms of different genotypes associated with sex. This explanation posits that the environment does not play an important role in the differences at hand, since the differences remain constant independently of the time and place of the observation.

The second scientist, an anthropologist, proposes a presumably rival explanation of the data. The anthropologist notices that height has increased throughout the centuries, thus suggesting that something must have happened as time went by that allowed people overall to grow more. Furthermore, this must be independent of factors such as sex, since both males and females have grown taller as time passed. The anthropologist then hypothesizes that this difference is due to changes in nutrition, which respond to changes in the environment and the development of technical artifacts that enabled better crops.

So, which of the two scientists better explains the data at hand? When we examine this situation, we see that the question is ill-posed. The reason is that the two scientists are interested in different phenomena altogether. The biologist is interested in differences between males and females, regardless of time; the anthropologist, on the other hand, is interested in differences across time independent of sex. This calls for a reconceptualization of the phenomenon to be explained. Initially, the two scientists thought they were explaining the data. After some disagreement and controversy, they realize that they are actually explaining different aspects of the same data, but not the same phenomenon. Hence, they redescribe what they wanted to explain, i.e., they reconceptualize the explanandum of each of their theories.

On this scenario, the both scientists have reinstated their explanations by shifting to different levels of analysis. Interestingly, the anthropologist has moved, not to a lower level, but to a higher one. What the anthropologist is interested in is not a lower level mechanism explaining differences in height, but rather more abstract interactions leading to those differences. As Kronfeldner puts it, the anthropologist is interested, not in a single difference, but in a “difference of differences."

If this assessment of the situation is correct, it follows that what the anthropologist has done is reconstitute the phenomenon by abstracting, rather than going into a lower level. Consequently, reconstitution works both ways. We can reconceptualize phenomena by either shifting to a lower level, as the biologist and the cases in genetics discussed by Bechtel and Richardson, or by shifting to a higher level, as the anthropologist does. How can we determine then where to go? How do we decide which level is the correct level for our explanandum?

5.3 How to carry out reconstitution

Consider first Kronfeldner's case of the anthropologist and the biologist with competing explanations and explananda. What leads the anthropologist to reconstitute their phenomenon by shifting to a higher level of analysis is, as Kronfeldner presents the case, the research question they intend to answer. It is because they are interested in answer the question "Why are people taller nowadays than they were in the fifteenth century?" that they shift to an abstract level of inquiry and reconceptualize their explanandum. So construed, reconstitution needs fixing a phenomenon that is already given, in this case, the differences in height at two points in time.

How do we fix the phenomenon in order to reconstitute it? We need some way of pointing to the phenomenon. In Kronfeldner's discussion, the anthropologist uses a piece of data: measurements of height at two points in time. This allows them to use the data, or a representation of the data (e.g. a graph), to make reference to the explanandum of their interest. In this sense, when discussing with the biologist, the anthropologist takes a step back and uses a device to ostensively shift to a different explanandum. Put differently, the anthropologist points to the phenomenon of their interest in order to clarify that they are not aiming at explaining the same phenomenon as the biologist. This justifies their reconceptualization of the phenomenon in a vocabulary that captures the explanations they are interested in.

We can construct a similar case for Bechtel and Richardson's story. On the face of conflicting hypotheses and results, reconstitution by shifting to a lower level of analysis occurs when scientists obtain a device to make reference to the phenomenon of their interest. In the development of genetics, scientists were interested in a range of observable phenomena. These include, first, differences between offspring of a given species (peas, *D. megalonaster*, or *N. crassa*). But second, and most importantly, they intend to explain the fact that certain properties are inherited independently of others, i.e., some form of independent assortment.

In a similar fashion to the anthropologist, we can imagine the following scenario. One geneticist claims that independent assortment is plainly false, since they had evidence that some observable traits did not pass on independent from each other (e.g. white eyes and sex in the *D. megalonaster*). Another geneticist, after Beadle and Tatum's results, might reply 'Independence assortment does obtain, but at the level of enzymes. It doesn't matter if it does not obtain at the level of observable traits, since differences in inheritance in general, such as *these* [pointing at the differences in how *N. crassa* thrive] are what I want to explain.' By pointing to the phenomenon of their interest, they can know look for the level of analysis that allows them to construct a theory and an explanation, that is, to reconstitute the phenomena.

In sum, reconstitution requires, as a first step, taking a step back and finding a device that allows scientists to make reference to the explanandum of their interest. In some cases, this may be the product of an experiment or a representation of obtained data. In any of these cases, these devices serve an ostensive function: they allow the

scientists to ostensively fix the explanandum and look for a level of analysis that allows them to cash it out and offer an explanation.[47]

Once researchers have fixed the explanandum, the next step towards successful reconstitution is picking out the level of analysis. This amounts to deciding which inferences are researchers interested in doing. Depending on the level of the inferences, researchers can then pick out the correct theoretical vocabulary to construct new concepts and hypotheses that will figure in the reconstituted phenomena.

Going back to our previous cases, the procedure can be spelled out as follows. In the case of the anthropologist, once they pick out their explanandum as the differences in height across time, they can specify which inferences they intend to make. In this case, these are inferences in terms of social factors such as culture and history as influencing nutrition. Hence, they will cash out the explanandum as a social phenomenon, rather than a merely biological one as the biologist does.

A similar case obtains in the scenario of genetics. Once researchers pick out their explanandum as differences in inheritance, they proceed to specify the level of inferences. In this case, they concede that they can make important inferences by going into a lower level of abstraction, to a biochemical level, as in this level they observe the phenomenon in question as well in the form of differences in producing enzymes. Since this level reinstitutes their capacity to decompose and localize mechanisms, this enables them to carry out inferences again and hence constitutes a possible route to reconstitution.

This step of identifying the level of inferences that scientists want to make has a crucial consequence for the project at hand. In the previous chapter, I claimed that the decision of which scientific kinds are involved in a given classification amounts to the decision of the level of inferences we want to make. Hence, this step in reconstitution leads to a selection of the type of kinds we want to construct. In other words, this step in reconstitution dictates the type of kind the explanandum phenomenon will figure in. As I will explain later, this is central to the case of emotions, as with this procedure we can make clear what type of kinds emotions constitute. In my view, this can be carried out by recognizing folk emotion concepts as ostensive devices that allow us to pick out the phenomena to be explained and carry out reconstitution.

Lastly, once scientists are clear on the explanandum phenomenon and the sort of inferences they intend to draw, the last step is finding the vocabulary to redescribe the explanandum and spell out their inferences of interest. This can occur by adopting the vocabulary of an already given scientific theory. This is the case of integrating the question of inheritance with the vocabulary of biochemistry. This might also be the case of the anthropologist in Kronfeldner's scenario. As I have presented it, the anthropologist might describe the phenomena they are interested in in terms of a social theory.

[47] This is not to say that all uses of representations in science are ostensive or that scientific concepts are formed through ostension. I am only claiming that when disputes about explananda happen, concepts and other representational devices in science may serve an ostensive function.

Given this account of reconstitution, we can now address the question of how a reconstitution of emotions could work out. Following the aforementioned steps, we must identify a way of fixing the explanandum, an ostensive device that allows us to make reference to the phenomenon we are interested in. Once we have such a device, then we can examine how to describe that phenomenon in a target scientific vocabulary.

5.4 Explication

The first step in reconstituting emotions is to identify the vocabulary to describe the explanandum phenomenon. In my view, this is the vocabulary of folk-psychology. Emotion concepts, first and foremost, are folk-psychological concepts. It is in our everyday interactions that we describe our behavior as that of fear, anger, sadness, happiness, etc. Hence, reconstituting emotions requires an analysis of our folk-psychological vocabulary. With this analysis at hand, we can then establish a theoretical vocabulary that captures the inferences we are interested in making and that do justice to the pretheoretical characterization of the explanandum.

This procedure of taking a given concept, specified by a set of pretheoretical tools (in this case folk-psychological concepts), and reworking it in terms of a scientific theory amounts to what philosophers of science call *explication*. This means that reconstituting emotions (and perhaps other folk-psychological phenomena) takes the form of explicating folk-psychological concepts. In what remains of this chapter, I will present an account of explication as applied to emotion concepts. This will set the stage for the next chapter, where I discuss the features of folk emotion concepts and how they can be applied in the construction of a scientific theory of emotions.

The traditional account of explication is due to Carnap. In one canonical formulation of explication, Carnap (1950/1963) presents it as follows:

> The task of *explication* consists in transforming a given more or less exact concept into an exact one or, rather, in replacing the first by the second. We call the given concept (or the term used for it) the *explicandum*, and the exact concept proposed to take the place of the first (or the term proposed for it) the *explicatum*. The explicandum may belong to everyday language or to a previous stage in the development of scientific language. The explicatum must be given by explicit rules for its use, for example, by a definition which incorporates it into a well-constructed system of scientific either logicomathematical or empirical concepts. (Carnap, 1950/1963, §2)

Carnap's classic example of an explication concerns the everyday concept of FISH in terms of the biological concept of fish or PISCIS. In everyday language, FISH includes organisms which, from the perspective of biology, do not qualify as fish, such as whales and seals. Hence, biology introduces a new concept, PISCIS, as an explicatum of FISH. PISCIS better captures the inferences that biologists make, hence providing a more fruitful classification than FISH.

According to Carnap, a successful explication is one that fulfills four criteria:

Similarity The explicatum must bear enough similarity to the explicandum so that the explicatum can be used in most cases where the explicandum has been used.

Exactness The explicatum must be characterized in exact terms in order to introduce it in a well-connected system of concepts.

Fruitfulness The explicatum must be fruitful in that it must be explanatorily useful.[48]

Simplicity The explicatum should be as simple as all other criteria permit. (adapted from Carnap, 1950/1963, §3)

It is central to highlight that in Carnap's construal, there is not one correct explicatum, but a number of them. The reason is that since the explicandum is relatively less exact than the explicatum, there will be a number of possible explicata that capture different ambiguities present in the explicandum. In the case of FISH and PISCIS, the second captures most instances of FISH but excludes whales. Other possible explicata may be constructed to capture whales, however, depending on the purpose of the explication. For example, if we were classifying animals by their habitats, an explicatum of fish may be constructed as "animal that lives in the ocean," hence including whales. Whether or not this explication is successful or not depends on whether it satisfies the purposes of the explication. In the case of biology, it certainly does not, given that biology draws inferences in terms of evolutionary history. Nevertheless, the success of an explication remains relative to a particular project. As Carnap explains:

> [...] if a solution for a problem of explication is proposed [an explicatum], we cannot decide in an exact way whether it is right or wrong. Strictly speaking, the question whether the solution is right or wrong makes no good sense because there is no clear-cut answer. The question should rather be whether the proposed solution is *satisfactory*, whether it is more satisfactory than another one, and the like. (Carnap, 1950/1963, §2, my emphasis)

Understanding the last step of reconstitution as explication has interesting consequences for our project. It allows us to restrict the possible paths on which we carry out reconstitution, and clarify how this procedure can lead to scientifically interesting concepts. By restricting reconstitution to successful explication, we get some criteria to evaluate whether the proposed reconstitution is satisfactory.

In this line, a successful reconstitution must be one that not only reinstitutes capacity to draw inferences, or to decompose and localize mechanisms, but one that

[48] Carnap's original claim is that the explicatum must be useful for the formulation of many universal statements. This commits explication to a limited version of science in which science only cares about the formulation of universal statements. We can however relax this criterion to better capture Carnap's intended meaning: that an explicatum must serve the explanatory purposes for which it is proposed.

leads to the construction of concepts that capture the explanandum phenomenon adequately (similarity condition), can be integrated into scientific theories (exactness condition), and that is ideally presented in the simplest form possible. Also, following Carnap's remarks on explication, some of these criteria are more useful than others. As in Carnap's account, the most important criteria are the fruitfulness and similarity conditions, while the others are desiderata that should be fulfilled only as best as possible.

Yet, Carnap's account of explication needs further clarifications if we are to apply it to emotions. The reason is that the methodology of explicating folk concepts carries with it an important tension that we have to bear in mind. This is the tension between maintaining reference to the target phenomenon, i.e., avoiding a change of subject, while offering a fruitful scientific construal. In what is left of this chapter, I will explore this tension and offer some remarks on how to best avoid its pitfalls.

5.4.1 Strawson contra Carnap

One influential argument against Carnap's account of explication was put forward by Strawson (1963). Strawson sets off by identifying the explication as part of the project of clarifying everyday concepts. He describes Carnapian explication as follows:

> A pre-scientific concept C is clarified in this sense [explicated] if it is *for certain purposes* replaced (or supplanted or succeeded) by a concept C' which is unlike C in being both *exact* and *fruitful*. The criterion of exactness is that the rules of use of the concept should be such as to give it a clear place 'in a well-connected system of scientific concepts.' The criterion of fruitfulness is that the concepts should be useful in the formulation of many logical theorems or empirical scientific laws. (Strawson, 1963, p. 504)

Strawson criticizes this account of explication by arguing that clarification, in the sense of introduction of scientific concepts in place of those in everyday life, is a change of subject. In his words:

> [...] however much or little the constructionist technique is the right means of getting an idea into shape for use in the formal or empirical sciences, it seems prima facie evident that to offer formal explanations of key terms of scientific theories to one who seeks philosophical illumination of essential concepts of non-scientific discourse, is to do something utterly irrelevant—is a sheer misunderstanding, like offering a text-book on physiology to someone who says (with a sigh) that he wished he understood the workings of the human heart. (Strawson, 1963, pp. 504-505)

As is clear from the quote above, Strawson thinks that because language has many uses, it is a mistake to think that the language of science could supplant all other uses besides explanation and prediction. In his view, either the operation of clarifying

everyday concepts in this way would be unfeasible, or it would yield concepts so different from the originals that we could no longer say that they are doing the same thing (i.e. change of subject).

Even though Strawson's argument is undoubtedly one of the most influential versions of this worry, he was not the only one to frame it, nor is it unique to Carnap's account of explication. In its broader construal, this worry has been called the *paradox of analysis*. Dutilh Novaes and Reck (2017) formulate it as follows:

> In its simplest form, the two premises of the paradox are: for an analysis to be correct, the analysans must be identical to the analysandum; but for the same analysis to be informative, the analysans must be somehow different from the analysandum. The conclusion is that no analysis can be both correct and informative: if an analysis is correct, it is not informative; if it is informative, it is not correct. (Dutilh Novaes & Reck, 2017, p. 212)

Besides formulating the paradox, Dutilh-Novaes and Reck discuss a solution. In their view, and in the spirit of Carnap's view, Dutilh-Novaes and Reck claim that we can reject the first premise of the argument, that in order for an analysis (or an explication) to be correct, it must be identical to the analysandum (or explicandum). There can be successful analyses or explications that lead to concepts that are not exactly identical to those they intend to clarify. Granted, there will be a mismatch between the two sides of the analysis, but we must be prepared to allow some degree of mismatch.

Consider again the case of FISH and PISCES. It is clear that these concepts are not identical with each other, as the first includes whales while the other does not. In this case, we have some degree of mismatch between the folk concept of FISH and its explicatum PISCES. Nevertheless, this mismatch need not imply that the explicatum does not offer a correct analysis of FISH for the purposes of biological classification. For biology, classification in terms of PISCES covers most of the instances covered by FISH while providing a more fruitful concept, as it allows inferences in terms of evolutionary history. Furthermore, FISH and PISCES still share extensions to a robust degree, and they can even be applied interchangeably in a number of contexts (as when referring to trouts or salmon).

As a result, some degree of mismatch between the folk concept and its explicatum is to be expected, given that scientific discourse differs from folk discourse. Yet, this does not imply that explication is doomed to failure. Mismatch does not need to constitute a change of subject. As long as there is still a robust degree of similarity between the explicandum and its explicatum, we can offer concepts that are not co-extensional but still refer to a great number of the same cases and that provide more fruitful conceptual tools with which we can construct scientific theories.

This line of reply has some central consequences for our project. First, it prepares us to accept some degree of mismatch between our folk concepts and the concepts we obtain after we carry out reconstitution and explication. This means that the explicata we propose for emotion categories need not be exactly identical with folk concepts.

The precise extent to which this mismatch can obtain, however, must be spelled out. I will tackle this problem below. For now, it suffices to say that some degree of mismatch is acceptable, and does not imply immediately a change of subject.

Second, the tension between these types of concepts can be used to shed light on another important aspect of reconstitution, at least in the context of folk vocabulary. Scientific discourse is not isolated from our folk concepts. We come in contact with scientific concepts in a number of ways. Since these scientific concepts differ in some respects from folk concepts, but are nonetheless in contact with them, we can expect interesting interactions between these two. In other words, we can expect scientific concepts to lead to changes in our folk concepts. If scientific concepts are fruitful so as to lead a progressive research program, it is plausible that they shift our folk concepts as science progresses.

This brings to light an important aspect of reconstitution and explication, namely, that they are continuous processes by which both science and prethereotical intuitions can change. As folk concepts change, we reconstitute and explicate the folk concepts into new scientific ones. And as science progresses, it generates conceptual change in the folk realm that eventually call for further reconstitution and explication.

Leaving this observation aside, let us pay attention to a second problem with Carnapian explication in the context of reconstitution. From Strawson's objection, we learned that we must bear in mind that there will be some mismatch between explicanda and explicata, but we must be careful not to propose concepts that end up in a change of subject. This worry is particularly present in an influential argument regarding the characterization of emotions as natural kinds, one proposed by Scarantino when characterizing different aspects of emotion research. Let us examine this argument in detail.

5.4.2 Scarantino's two emotion projects

Besides the worry that explication is inherently irrelevant since it risks a change of subject, there is also a worry coming from the opposite side, that is, a worry that making scientific concepts match too closely with folk concepts will prevent progress in scientific research. In the case of emotions, this worry is present in Scarantino's (2012) suggestion to keep folk emotion concepts and scientific concepts separate. I have mentioned Scarantino's worry above, but now I will discuss it in detail.

Scarantino's worry stems from a discussion of traditional accounts of emotions. In his view, traditional accounts have assumed that we ought to map emotions onto some homogeneous theoretical construct, whether in terms of neurobiological mechanisms, social constructions, etc. In any of these cases, for each of the main attempts to carry out such a mapping, there are a number of counterexamples that seem to falsify the candidate scientific construction of emotions. This leads to what he calls the *problem of scope*:

> **Problem of Scope** For every scientific theory T that tells us that an emotion/an-ger/fear/ etc. is X, we can find counterexamples con-

> sisting of things called "emotion"/"anger"/ "fear"/etc. in English that are not X, and/or things not called "emotion"/"anger"/ "fear"/etc. in English that are X. (Adapted from Scarantino, 2012, p. 361).

For Scarantino, the assumption leading to the problem of scope is that we must find some unity in traditional emotion categories. In other words, Scarantino takes issue with the claim that folk-psychological concepts themselves refer to a homogeneous set of phenomena and thus demand a homogeneous theoretical construction if they are to be integrated in a scientific research agenda.

To reject this assumption, Scarantino distinguishes what he calls the *Folk Emotion Project* and the *Scientific Emotion Project*:

Folk Emotion Project Offer a *descriptive* definition of the conditions of membership of traditional emotion categories such as emotion, anger, and so on.

Scientific Emotion Project Offer a *prescriptive* definition of the conditions of membership of natural kinds of emotion, natural kinds of anger, and so on. (Adapted from Scarantino, 2012, p. 364).

On this construal, the Folk Emotion Project has only descriptive aspirations, as it is only the project of describing how speakers use emotion concepts. The Scientific Emotion Project, however, does away with how speakers use emotion concepts and focuses on the correspondence between emotions and natural kinds.[49] In Scarantino's view, we must be prepared to accept a plurality of natural kinds for each emotion category (both for emotion as a general category as well as particular emotion categories). For all we know, "anger" as a folk-psychological term might map onto many natural kinds of anger, each demanding its own description by means of different theoretical frameworks.

Presented in this way, Scarantino's account separates folk emotion terms from scientific categories in order to avoid problems with mapping the first one-to-one onto the latter. Nevertheless, without further constraints, making this excision between these two vocabularies leads directly to the risk of changing the subject that Strawson worried about and that I have invoked at several points throughout this work. The reason is that whatever kinds we find for emotions, we need to have good reason to call them kinds *of* emotion. Suppose we find a kind that presumably is a kind of anger. Whatever this kind is, there must be criteria to call it a kind of the phenomena we call "anger."

These problems bring out a tension when it comes to reconstitution and explication. When we reconstitute the phenomena, we pick out the phenomenon we are interested in and proceed to construct a scientific vocabulary to describe it. The resulting scientific construct must overlap to some degree with our pretheoretical conception of the phenomenon, otherwise leading to a change of subject. However, making this

[49] I shall use the term "natural kind" to make justice to Scarantino's proposal, in spite of my skepticism about the term presented in the previous chapter. I will present some clarifications of this use below.

overlap too demanding can lead to assumptions that stagnate scientific progress, as in the case of assuming that emotions must map one-to-one onto natural kinds. Consequently, there must be some mismatch between pretheoretical and scientific constructs (as with the Folk Emotion Project and the Scientific Emotion Project) but it must be a constrained mismatch so as not too lose track of the phenomenon of interest.

Scarantino offers two constraints to solve this tension between folk and scientific categories. In his words:

> A good prescriptive definition of emotion/anger/fear/etc. should specify the condition of membership of a natural kind of emotion/anger/fear/etc., namely a transformed category provisionally called "K" such that (a) K's members are the maximal class of items that tend to reliably share inductively and explanatorily important properties on account of one or more causal mechanisms (naturalness condition), (b) most or all of K's members are members of the traditional emotion categories of emotion/anger/fear/etcetera (similarity condition). (Scarantino, 2012, p. 366)

According to Scarantino then, a given construct candidate to offer a scientific description of an emotion must be such that the construct specifies a maximal class of objects that share a cluster of properties in virtue of a mechanism (i.e., form a natural kind according to Boyd's HPC account of kindhood), and whose members are, to some degree, members of the traditional emotion category that the kind is supposed to be a kind of. As I explained in the previous chapter, I do not think that Boyd's HPC account is the only way to construe scientific kinds. Nevertheless, let us follow Scarantino in this use for the time being.

Figure 5.1 displays different possibilities for mappings between these kinds and folk emotion term. Let E be a folk emotion term such as "fear" or "anger," and let $K_1 ... K_6$ be natural kinds in Scarantino's sense. On Scarantino's view, E involves two or three natural kinds. K_1 and K_2 are members of E because all of their instances are also instances of E. K_3 is also arguably a kind of E, since a vast majority of its members are members of E, even if some are not. In turn, K_5 and K_6 illustrate cases of kinds that do not belong to E because few or none of their members belong to the general category, respectively.

Scarantino's proposal does offer some clues in the right direction. In order to avoid changing the subject, there must be some robust degree of overlap between our emotion categories and the natural kinds, or in my case scientific kinds, corresponding to them in a given scientific theory. Yet, Scarantino's criteria require further development. The proposed criterion that most instances of an emotion kind must correspond to instances of the folk emotion concept requires clarification, as there may be cases where membership to the folk emotion concept might be ambiguous.

The contrast between K_3 and K_4 displays the problem with the proposal. Suppose that in the case of K_4, half of its members are members of E while the other half are not. Is K_4 a kind of E? If so, would it follow that in order for a kind to count as a member of a category such as E, at least half of its members must be members of

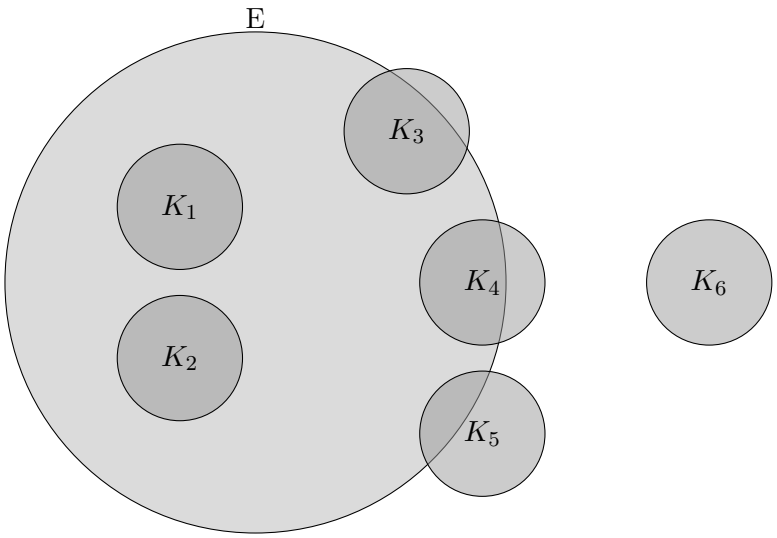


Figure 5.1. Scarantino's view of the relation between folk emotion terms and natural kinds of emotions.
E designates an emotion term such as "fear" or "anger." $K_1...K_n$, in turn, refer to natural kinds of emotion which may or may not be instances of E.

E? If not, why is K_3 a kind of E while K_4 is not? As Scarantino has formulated the constraint, it would be presumably because in the first case, most cases belong to E whereas in the other they do not. But where should we draw the line? In sum, how are we to decide whether a kind counts as a kind of E or not?

This ambiguity opens up an important problem and a crucial decision to be made when theorizing about emotions. In the spirit of Carnap's account—which is arguably the inspiration of Scarantino's constraints—one option is to count K_4 as a kind in case doing so provides an advantage when it comes to grounding inductive inferences and explanations, that is, in case it satisfies the fruitfulness criterion better than not counting it as a kind of E. But then, how are we to decide when does the inclusion of K_4 is more fruitful than its exclusion?

Scarantino's account depends on the underlying mechanisms producing the property clusters $K_1...K_6$. This makes the decision lie in the clusters themselves, not in their fruitfulness for scientific theories. As I have constructed the case, however, the decision cannot be made merely in terms of mechanisms, as these make it so that only half (or even half plus one) of the instances of K_4 are members of E. What we need, in my view, is to relax these constraints.

One way to make progress in this direction is to appeal to different types of kinds as I argued in the previous chapter. By appealing to only a specific type of scientific kinds (namely, HPC kinds), Scarantino restricts the possible inductive and explanatory patterns that can account for emotions. If my view on scientific kinds is correct, we can relax this criterion to accommodate other types of scientific kinds beyond HPC kinds. According to my account, a kind K_n counts as a kind of E if its members share inductively and explanatorily relevant properties in terms of what makes these kinds

projectible. By appealing to their projectibility, we introduce helpful concepts that enable us to clarify the situation above.

An advantage of a pluralistic account of kinds is that it allows comparing how strict or liberal a given taxonomy can be depending on the criterion on which it construes its kinds. For example, on an essentialist construal, a kind might be more restrictive than on a historical construal, since under the essentialist framework, all members of a given kind must share exactly the same concrete microstructural properties, whereas the historical framework appeals to causal histories which allow more variation across the kind. In some cases, such variability is desired, as in the case of species. By construing species historically, we make more categories that work better for the projections that biologists are interested in making than by appealing to an essentialist model. Hence, if the historical construal leads to a more projectible taxonomy, then that construal ought to be preferred. Put differently, we can cash out the fruitfulness criterion in terms of which model helps scientists make better projections. In the case above, historical kinds are more fruitful in this regard.

This offers one way to resolve problems about whether a kind is a member of an emotion category. Whether or not K_4 is a member of E or not depends on how well including K_4 in E helps scientists make projections across E. If we can find a framework to construe more powerful scientific kinds among the ones presented in the previous chapter (essentialist, historical, functional, and social), then membership to E will depend on whether that framework would lead to the inclusion or exclusion of K_4. For instance, it is quite likely that an essentialist construal would exclude K_4 from E, given that not all instances of K_4 share the same properties as other instances of E. In turn, it is plausible to construe E in terms of a social or functional kind that includes K_4 in spite differences in microstructures. Depending on which framework leads to more projectible categories, we will have a criterion to include or exclude the kind, thus disambiguating this situation.

This appeal to projectibility introduces an important meta-criterion to choose among different kinds and hence to disambiguate the case above, namely, entrenchment. As I explained in the previous chapter, whether or not a kind is projectible partly depends on whether it is well entrenched, that is, how successful its use has been in past projections. Resolving whether or not the inclusion of a kind in an emotion category is more or less fruitful must thus involve questions about how we have characterized that category and that kind in the past. This would give us clues about how well entrenched a given characterization is and thereby which taxonomy leads to better projections.

In this sense, the decision of whether to include K_4 is at least partially pragmatic. If all possible frameworks to construe scientific kinds of emotion lead to equally projectible categories, then we can look at past inferences to see which one fits best with past successes, i.e., which one is more entrenched. This is an interesting consequence for two reasons. First, it makes the situation potentially resolvable even if the choice between different construals of scientific kinds cannot decide the matter. It would be

up to emotion researchers to pick the kinds that best fit their previous explanatory practices.

This leads to the second upshot of this proposal, namely, that it gives space to considerations about how scientists actually use emotion categories and to the pragmatics of scientific research and the history of emotion science. In this account, explicating emotions is not a task done over and above actual scientific practice, but an integrated part of it. In order to explicate emotions, we must not only work towards projectible categories, but we must do so considering past projections that have been successful in advancing emotion research.

In sum, I concede to Scarantino that a robust number of instances of the kinds we identify with emotion categories must also be members of the folk category. In ambiguous cases, the decision to count a kind as a given emotion kind, however, must depend partly on pragmatic criteria appealing to projectibility and fruitfulness. To make this clear, I will synthesize the previous discussion under the following proposal.

5.5 Conclusion: How to explicate emotions

To close this chapter, I will sketch a strategy to explicate folk emotion concepts and create new, fruitful scientific concepts. I will do this by following what I take are the main lessons we can learn from the different discussions presented above, namely, that on reconstitution, explication, and the tension between folk and scientific concepts of emotion. In the next chapter, I will apply this strategy to defend a functionalist model of emotions.

Recall first the two-step account of reconstitution I proposed above. According to this account, we reconstitute the phenomena by (1) having some ostensive device that allows us to point to the phenomenon of interest; and (2) picking out the level of analysis with which we will construct scientific categories to study that phenomenon, i.e., decide which framework of scientific kinds fits best to our explanatory purposes.

In the case of emotions, as I claimed above, the first step, finding an ostensive device to point to the phenomena of interest, can be done by analyzing folk emotion concepts. Folk emotion concepts provide the first approximation to what emotions are, hence enabling us to detect what the explanandum of a scientific theory of emotions is.

Given the appeal to folk emotion concepts as a first step in the construction of a scientific theory of emotions, the second step, finding the correct level of analysis, amounts to explicating emotion concepts in terms of one of the aforementioned frameworks of emotion kinds. The choice of framework will depend on how well each framework captures properties that characterize emotions in our folk psychology while allowing the formulation of projectible kinds. In other words, we must pick the best framework in terms of how well they mirror the properties emotions have in our folk psychology.

In my view, this procedure will allow us to abstract away from folk terms into scientifically tractable ones. Instead of maintaining folk concepts as they are and

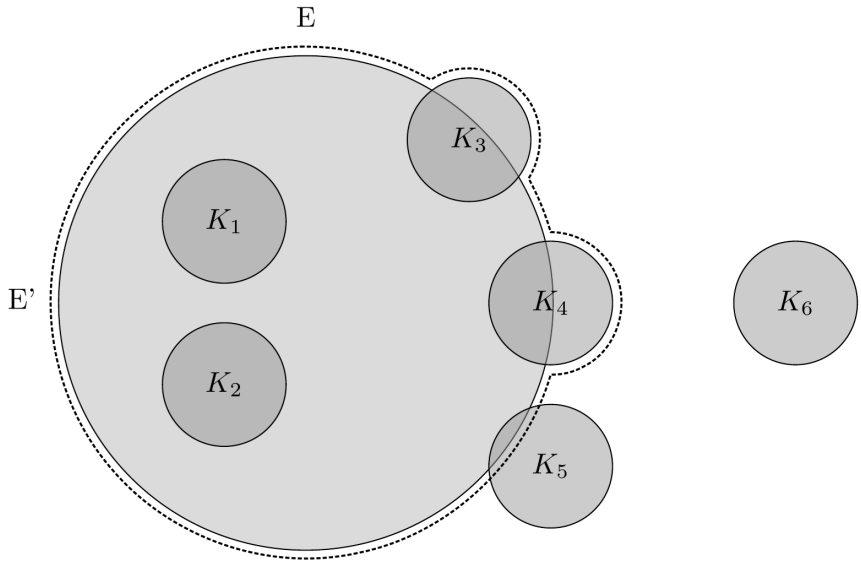


Figure 5.2. Sketch of the difference between folk and scientific emotion concepts. As in Figure 5.1, E designates a folk emotion concept and $K_1...K_n$ candidate natural kinds for emotions. In this case, we now see E' (delimited by the dotted line) as a scientific emotion concept for E, which does not map one-to-one onto each case of E (hence allowing some mismatch) but which overlaps robustly with E.

attempting to use them in science, this procedure yields new concepts that are robustly similar to the original ones (meeting the similarity criterion) while being scientifically fruitful.

Figure 5.2 displays the result of this reconstitution procedure for a given emotion concept. Just as in the case of Scarantino's proposal, E represents a folk emotion concept, and $K_1...K_6$ represent property clusters that are either associated or not with that emotion concept. These can be neural, physiological, or behavioral property clusters (or a combination thereof). In some cases (K_1 and K_2), all of the members of the kind are members of E. In other cases ($K_3.K_4$, and K_5) only some members of the kind are members of E. In this picture, rather than changing E to make it fit these kinds, we can construct a new scientific concept E' which will include the kinds that it requires to make a scientifically tractable, projectible kind. In this scenario, E' includes the members of kinds K_3 and K_4, but not members of K_5.

As we can see in the figure, there is some degree of mismatch between E and E'. This mismatch prevents us from overfitting scientific categories to the folk concept in question, hence allowing some freedom for science to determine whether certain instances will count as members of an emotion category depending on whether the inclusion or exclusion of these instances make the category projectible. This helps meet Scarantino's worry that by forcing science into using folk categories, we stagnate scientific progress.

Additionally, the account sketched here also prevents changing the subject, as it is clear how E' is constructed. As I have explained above, E' is constructed by analyzing the folk concept E and, when it comes to borderline cases, evaluating whether or not

the inclusion of marginal kinds K_3, K_4, and K_5 makes a more fruitful category. This makes it clear what the connection is between the scientific concept E' and its folk relative E.

One interesting consequence of this account of the reconstitution and explication of folk emotion concepts is that it renders the mismatch between these concepts and their scientific counterparts productive. Since scientific concepts need not have exactly the same extension as folk concepts, but they maintain a robust degree of similarity, they can push folk concepts into new directions. This allows for an account were scientific findings inform folk concepts and lead to conceptual change in line with scientific progress. Additionally, if there is such conceptual change in folk concepts, we can also expect scientific theories to be reworked to accommodate new pretheoretical intuitions and maintain in sight the phenomena of interest. Hence, the account I sketch here provides an image where folk psychology and scientific theories are not in opposition, but work together in a continuous fashion, while avoiding collapsing scientific categories into folk taxonomies.

In the next chapter, I will apply this strategy to reconstitute emotions and answer the question of which framework to construct scientific kinds works best for emotion research as it currently stands. If this strategy is successful, we have good reasons to be optimistic that the Theoretical Challenge can be overcome. In turn, this will shed light on how to approach the Empirical Challenge as well, thus making an important step towards the construction of a scientifically respectable theory of emotions.

Chapter 6

A Functionalist Approach to Emotion Kinds

In this chapter, I explore the issue of how we can successfully reconstitute emotions for the purposes of scientific investigation. I will follow the procedures sketched in the previous chapter (chapter 5), and integrate them into the picture about kinds developed in the chapter before that (chapter 4). This procedure calls for an analysis of our pretheoretical commitments in the form of diagnostic features that help us identify the phenomenon we want to construct as a scientific kind. In the case of emotions, I will examine some features that are present in our folk-psychological vocabulary. Afterwards, I will approach the question about kinds by discussing each type of kind as presented before, and I will argue that among the types discussed, functional kinds are the most suited to construct scientific concepts of emotions. In the rest of the chapter, I will defend this view against some objections and explore some of the advantages of this account.

To be clear, this chapter does not propose a theory of emotions in itself. These observations are intended as meta-theoretical observations, that is, as regarding how we can conceive and construct candidates to successful theories of emotions. These remarks do have important consequences for specific theories emotions, as it allows us to cast doubt on current theories and offer support for others. Nevertheless, my positive view focuses on the conditions for success, rather than a particular version of the view I am proposing. In other words, I am offering an account of how to develop scientifically interesting concepts and theories of emotions, while leaving the specifics of what these concepts and theories are to scientists.

6.1 The Folk-Psychology of Emotions

Many studies on emotions have investigated what we can call the folk-psychological picture of emotions. These studies pertain to what Scarantino called the *Folk Emotion Project*. While many of these studies are not intended primarily as studies on folk-psychology, they do provide interesting sources of information as to how folk-psychological concepts of emotions work. Thus, since we are interested in how the

folk-psychology of emotions makes reference to diagnostic features, I will pay special attention to studies on emotion concepts and emotion attribution.

Additionally, some of the features I shall list below will be included by appeal to facts about how we talk about emotions more generally. By analyzing evidence and arguments in these directions, we can make ourselves an idea of the properties that emotions have in the folk-psychological domain, which we will later try to capture in terms of scientific concepts. Hence, in this section, I will go over empirical evidence that can be put to use to tap into facts about our folk-psychological account of emotions, and that can provide a stepping stone to carry out an empirically informed form of conceptual analysis later on.

6.1.1 (Relatively) Uncontroversial features

Many researchers agree on some of the features emotions exhibit in our folk-psychology. Among these, perhaps the most prominent features are *valence* and *arousal*. Valence refers to the positivity or negativity of each emotion. Arousal, in turn, refers to the level of excitation or inhibition that we feel when having an emotion. For example, happiness is a positive emotion that excites us; sadness, a negative emotion that leaves us unenergetic. While I do not think that emotions are reduced to valence and arousal (see chapter 1 on constructionism), these are still important features present in our folk-psychology.

Evidence for the importance of valence and arousal in how we conceptualize emotions pretheoretically can be found in studies on dimensional models of emotions. One such early model was Russell's circumplex model of affect. Russell (1980) conducted a number of studies to develop this model. In one study, subjects mapped descriptions of emotions onto emotion categories. These descriptions made reference to whether an emotion was arousing or not, pleasurable or unpleasurable, distressing or contentful, and exciting or depressing. In another study, subjects rated how accurate adjectives in a list were with regards to a set of emotions. These studies showed that the main dimensions that accounted for the observed variance were arousal and pleasure (i.e., valence). This means that emotions can be described in these terms to some extent. Later studies would sophisticate this model, but the mapping of emotions onto valence and arousal remained (Remington, Fabrigar, & Visser, 2000; Russell, Lewicka, & Niit, 1989; Yik, Russell, & Steiger, 2011).[50]

Besides emotions being valenced and arousing states, researchers also agree that emotions motivate actions. All of the theories of emotions discussed in chapter 1 share this common theme. For basic emotion theories, emotions are related to actions regarding survival; for appraisal theories, emotions are syndromes leading to actions according to an appraisal; for constructionists, emotions are acts of categorization

[50] This interpretation is much weaker than Russell's. Russell takes these studies to support the view that emotions have not only this conceptual structure, but that they are constituted by these dimensions. This will later become the basis for the construct of *core affect*. I do not endorse this interpretation for the arguments offered in previous chapters (see chapter 1 on constructionism). Yet, the weaker interpretation allows us to make use of these results in terms of what they tell us about the folk-psychological structure of emotion concepts.

that can bring about actions. This recurrent theme stems, I believe, from the fact that in our folk-psychology, emotions are thought of as closely related to motivation.[51] Empirically speaking, there is evidence to support this association between emotion and action. Much of the literature explored in sections §2.2.3 and §2.2.5 suggest that our emotion concepts are linked to action tendencies such as wanting to escape, fight, flee, attack, approach, and the like.

There is also evidence that emotions are considered to be embodied by our folk-psychology. By embodiment, I mean that they involve physiological reactions. Scherer and Wallbott (1994) provide interesting evidence in this direction. They conducted a study in which subjects answered open-ended questions about how they describe their emotions. Importantly, this study included several cultures, including European, Asian, American, and African countries. Overall, subjects described their emotions as involving physiological symptoms such as breathing changes, muscles tensing, stomach trouble, and differences in felt temperature. This means that across different cultures, folk emotion concepts are thought of as involving bodily reactions.

In a recent study, Hietanen, Glerean, Hari, and Nummenmaa (2016) showed a similar result in how children learn emotion concepts. In this study, the researchers investigated the development of bodily sensations associated with basic emotions in 6 to 17 year olds. They used a paper-and-pencil version of emBODY, a tool to map emotion categories to bodily activity. Participants were given an A4-sized piece of paper showing two outlines of a human body and an emotion word between them. Participants read the emotion words and colored the bodily regions whose action typically felt becoming strong and faster on the left body or weaker and slower on the right body when feeling each emotion.

The researchers found that as a child's emotion concepts become more refined, they exhibit associations between each emotion category and bodily features. For example, happiness is thought of as involving activity in the upper part of the body, whereas sadness involves a general feeling of inactivity. These results suggest that in our folk-psychology, the body plays an important role in how we distinguish between different emotions.

Another important feature of emotions in the folk-psychological realm is their appeal to phenomenality. It is common to describe our emotions as *feelings* of a certain kind. This may be associated with feelings in the body, as shown by studies such as the one by Hietanen et al. (2016) mentioned before. They can also be associated with experiential states more generally, as seen in phrases we use when we say things like "I feel sad" without appealing to a specific body part. It is worth noticing that, however, as Ortony, Clore, and Foss (1987) remark, emotional vocabulary usually accepts both constructions with the verb 'to feel' and the verb 'to be'. This is to say, for emotional words, we can apply them in sentences such as "I feel afraid" and "I am afraid" without changing their intended meaning.

[51] Whether or not this relation is one of causation or not, I will not address at the moment. See Scarantino (2017) for a discussion.

The fact that we can apply both constructions helps distinguish emotions from other related phenomena in our folk-psychological vocabulary. For example, using both constructions without a change in meaning does not obtain for other words, such as 'abandonment,' to use Ortony and colleagues' example. When we say "I feel abandoned," we may be expressing that we feel certain emotions provoked by the belief that we are abandoned. In my view, we could even be implying that we believe we are abandoned, although we do not have conclusive evidence to back up our claim. In contrast, when we say "I am abandoned," we are expressing a state of affairs or a fact, i.e., we are asserting that we are indeed abandoned.

This linguistic fact about folk emotion concepts highlights the importance of feeling in how we think about emotions. Later I will argue that this should not be confused with the claim that emotions are to be understood as mere feelings, nor does this imply that we must include phenomenality as a central piece in the construction of a theory of emotions. For now, let it be noted only that the appeal to feelings and phenomenality is a feature of emotions as understood by our folk-psychology.

One last relatively uncontroversial feature of emotions is their intentionality. When we claim that we or someone is experiencing an emotion, we say that the emotion is about some object or event. Intentionality is clearly seen in phrases such as "I am angry that..." or "I am sad about...," where the ellipsis can be replaced by a proposition or an object or event, respectively. Few studies have addressed intentionality explicitly though. Nevertheless, it is worth including in the list of features that help us point to the phenomena we call emotions.

The features discussed above offer at least an approximation to the phenomenon a theory of emotions must explain. It is worth clarifying that these do not constitute necessary or sufficient properties for a state to qualify as an emotion. As I explained before, they only constitute heuristic features that help us make reference to the explanandum phenomenon. Before I discuss how different models can accommodate these features, however, I would like to add two features which, I believe, are of importance but often neglected in discussions about emotions. These are the sensitivity of emotion attribution to contextual factors, and cases of attribution of emotions to non-human animals.

6.1.2 Context-sensitivity

In section §2.2.4, I mentioned some evidence that cultural factors influence how and which emotions we attribute to others. Here I will add evidence for context-sensitivity effects, focusing in how contextual cues aid in individuating emotions. Evidence in this direction can be divided in two groups. On one hand, we may ask how we identify emotions in ourselves, and how our context helps us make this identification. On the other hand, there is the question of how context affects our attribution of emotions to others.

Regarding the first question, one canonical experiment is that by Schachter and Singer (1962). Subjects in this experiment were told that they were in a study on the effects of vitamin supplements on vision. The experimenters told the participants that

they would be given a small injection of 'Suproxin', a fake drug which would affect their visual skills. In reality, subjects were given either epinephrine (i.e., adrenaline) or placebo. Those that received epinephrine were divided into three groups. The first group was informed of the effects of the epinephrine, such as accelerated heart rate, pupil dilation, etc. The second one told that their feet would go numb, some parts of their body would itch, and that they might get a headache. This was meant misinform subjects about their own states to avoid the information given from biasing subjective reports. The last one was told nothing about the effects of the injection.

After being injected with either epinephrine or placebo, subjects were taken into a room where they were asked to wait for 20 minutes while the 'Suproxin' took effect. Afterwards, they would be asked to do some vision tests. While waiting in the room, the experimenters would bring in a stooge, introduced as another subject. The stooge and the room were controlled to induce a given emotion. On one condition, the room would be slightly disorganized, but the experimenter would apologize, invite the participant to help themselves to scratch paper, rubber bands, and pencils, and the stooge would play around in the room. This was meant to induce euphoria. On the other condition, the experimenter would give the participant and the stooge five-pages long questionnaires designed to ask personal and insulting information. During this task, the stooge would complain about the shots and make annoying remarks about the questionnaire. This was meant to induce anger.[52]

According to Schachter and Singer, the idea was to investigate the role of cognition in emotion. In their view, if participants were given an explanation for their bodily symptoms, they would not think of these reactions as part of an emotional state. In other words, if they knew about the effects of epinephrine, any subsequent reaction would be attributed to the drug. However, if they did not have such information, subjects would think that they were in an emotional state. This state, Schachter and Singer hypothesized, would depend on contextual cues leading to a specific subjective interpretation of the situation (which is what Schachter and Singer mean by 'cognition'; see Reisenzein, 1983). In their words, "Given a state of physiological arousal for which an individual has no immediate explanation, he will "label" this state and describe his feelings in terms of the cognitions available to him" (Schachter & Singer, 1962, p. 381). To manipulate these cognitions or interpretations, the researchers fabricated contexts meant to lead the subject to judge their state as either euphoria or anger. As a result, they created a situation in which, presumably, two participants in the same physiological state of epinephrine-induced arousal may have had different emotional states depending on the context.

[52] There is a problem in the distribution of subjects into each of these conditions. While the experiment calls for a 4x2 design (Epi. Informed, Epi. Ignorant, Epi. Misinformed, Placebo vs. Euphoria, Anger), the Epi. Misinformed plus Anger condition was not included. The researchers argue that this is because they thought the Epi. Misinformed plus Euphoria condition would suffice to account for the effects of information. However, this introduces a problem: if there are differences within the Euphoria condition, they cannot rule out that the effects are specific to Euphoria.

Schachter and Singer report that in the epinephrine conditions, subjects showed significant increased pulse rate, palpitation, and tremor. When comparing between the epinephrine-ignorant and the epinephrine-misinformed conditions in the context of euphoria, subjects reported similar levels of euphoria. In both cases, they reported being more euphoric than in the epinephrine-informed condition. This suggests that in both cases, subjects relied on the context to identify their emotion and not on their physiological state, consistent with the investigators' hypotheses.[53, 54]

More recent evidence supports the claim that context affects how we attribute emotions to ourselves. Sabini and Silver (2005), for instance, hold that emotions are attributed in a three-sided context: the person having the emotion, the person describing the emotion, and the audience for the description. They claim that speakers decide how to describe emotional experiences based on information in all of these sources. In other words, it is pragmatics that decides how an experience is described and which emotion concept is appropriate. Consequently, distinctions between emotions are traced as a function of the context, rather than appraisals (as in appraisal theories) or mere experience.

To argue for their claim, Sabini and Silver present a series of examples of context-sensitivity effects. Two of the more salient ones concern the distinctions between envy and anger, and between shame and embarrassment. Regarding the first, they suggest that envy occurs when "a person recognizes that another's accomplishment make [sic] him or her look bad to self [sic] and (or) others" (Sabini & Silver, 2005, p. 4). If the person that looks bad makes an unwarranted accusation against an accomplished person, the first will be seen as envious. Nevertheless, from a first-person perspective, the envious person feels anger, and they are attributed envy only insofar as others know that their anger is unwarranted. In other words, anger and envy are distinguished from each other only in that envy depends on contextual factors which include someone's accomplishment and another's unwarranted anger towards that accomplishment.

Regarding the case of shame and embarrassment, the authors claim that these emotions are differentiated in terms of the quality of a perceived flaw in the self. Let

[53] For the anger condition, subjects did not report feelings angry or irritated. According to Schachter and Singer, only after de-briefing did subjects confess being annoyed by the questionnaire. This makes it problematic to interpret the results, since it is unclear to which extent did the manipulation work. Yet, the researchers report that subjects in the epinephrine-informed condition were less angry (i.e. happier, as they used the same happiness measure in both emotion contexts) than in the epinephrine ignorant and placebo conditions. Despite the problems with these results, the researchers take them as supporting their hypotheses.

[54] As influential as it has been, the Schachter and Singer experiment is problematic. Reisenzein (1983) explores some criticism of this experiment. He mentions a number of studies that attempted to replicate the results without success. For example, Marshall and Zimbardo (1979) tried unsuccessfully to replicate the euphoria condition following Schachter and Singer's design. They included several dosages of epinephrine, and found that subjects that had received the drug showed a tendency towards negative feelings rather than positive ones. Yet, other studies with similar paradigms on other emotions find some effects, albeit weak or indirect. For instance, in Cooper, Zanna, and Taves (1978), subjects who had been administered amphetamine and were asked to write a counterattitudinal essay subsequently change their attitudes in the direction implied by their behavior in contrast to subjects who had been given a placebo. The authors infer that cognitive dissonance, taken as an aversive state similar to an emotional one, had been intensified by the drug-induced arousal.

us elaborate. According to a study by Sabini, Garvey, and Hall (2001), the previous consensus that shame occurs when there is a violation of a moral norm, while embarrassment corresponds to the violation of a social convention, is misguided. Sabini et al. argue that (1) moral transgressions are not the only causes of shame, and (2) people can feel embarrassed when they are clumsy rather than violating social conventions. In their view, these emotions are distinguished in that in shame, agents believe that an event has revealed a real flaw of their self (moral or otherwise), whereas in embarrassment, the event has appeared to reveal a real flaw but it has really not. If the audience believes there is a real flaw, the reaction tends towards anger rather than embarrassment.

To test this, the researchers asked participants in an experiment to imagine a scenario where a colleague helps them move from one office to another. While moving, their colleague discovers some pornography among their stuff. In one condition, the pornography belonged to them. In another condition, the pornography belonged to the former occupant of the office, even though their colleague thought it belonged to them. In the first case, participants reported feeling both ashamed and embarrassed, while in the second, they reported feeling more embarrassed than ashamed. This result suggests that the difference between instances of shame and embarrassment lies in contextual factors. In this experiment, whether the participant felt shame or embarrassment depended on whether they owned the pornography (leading to a situation showing a fault in their selves) or not (leading to a situation where the fault is not in their selves, although it would appear as it were). Further studies on the physiology and expression of these emotions support Sabini and Silver's hypotheses (see e.g. Crozier, 2014).

Context-sensitivity effects can also be seen in cases where we attribute emotions to others. Conway and Bekerian (1987) tested how knowledge about a given situation could help subjects infer an ensuing emotion. They gave participants a list of sentences describing a situation, and asked them to decide which emotion might be associated with it or which emotion might someone feel in that context. They found that for all emotions, subjects named the same emotions equally homogeneously. In other words, subjects agreed significantly on which emotions corresponded to which situations. This suggests that contextual information plays at least a heuristic role in identifying emotions.

Ngo and Isaacowitz (2015) also studied the role of context in identifying emotional expressions. They presented subjects with a set of faces paired with a context, and asked subjects to specify the corresponding emotions while ignoring the context. In spite of this instruction, the researchers found that subjects were still influenced by contextual cues of the scene. Specifically, they report that subjects would often report the emotion corresponding to the context even if it was not supposed to match the given expression. If subjects saw a fearful expression in a context that would normally induced anger, they were prone to report the corresponding emotion as one of anger rather than fear, for example. In a follow-up study, they replicated these findings even when comparing participants who were told that the context was irrelevant with participants who were told the opposite. This indicates that regardless of the instruc-

tion, subjects would inform their decision based on the scene and not on the emotional expression alone. Similar findings can be found in Hareli, Elkabetz, and Hess (2018).[55]

All in all, there seems to be evidence that emotions are context-sensitive, at least to some degree. That is to say, how an emotion is identified and individuated depends at least partially to factors of the context. This is true regardless of whether people are told to ignore contextual cues, suggesting that this phenomenon might be quite automatic. This will become important later on, since in my view, essentialist and historical models have problems accommodating this feature.

6.1.3 Minimal attribution and non-human animals

In closing this section, I want to discuss the question: what are the minimal criteria on which we attribute emotions to others? It is uncontroversial that emotions exhibit various degrees of complexity, ranging from basic responses such as fight-or-flight ones in the case of fear, to sophisticated instances such as embarrassment due to failing to meet a cultural norm. To identify what is common to all of these instances, it is useful to think about the most simple cases with the hypothesis that whatever properties are present there can be extrapolated to more complex ones. Put differently, we can think of the minimal cases as providing a basis and later cash out more sophisticated cases as additions to that minimal case. I believe that attribution of emotion to non-human animals is an interesting minimal case. If at least some animals are said to have emotional reactions, this provides a case where we can put aside social and cultural factors involved in our folk-psychology of emotions.

To what extent to we attribute emotions to non-human animals? Often we describe animal behavior as emotional, such as when we say that our dog is happy to see us or when we say that our cat is angry because we have bothered it. To make these attributions, there must be something in the animal's behavior that leads us to apply emotion concepts to it. What properties then does non-human animal behavior exhibit such that we attribute emotions, at least in some cases?

In raising these questions, I draw from one of the most influential emotion researchers that focused on non-human animals: Darwin. Darwin thought that an investigation on the nature and mechanisms behind human emotional expression must include animals as well. The reason is that by looking at the similarities between human and animal expression, we can generalize to obtain general principles that explain how these expressions come about and how they allow organisms to signal each other. In his words:

> [...] I have attended, as closely as I could, to the expression of the several passions in some of the commoner animals; and this, I believe to be of paramount importance, not of course for deciding how far in man certain expressions are characteristic of certain states of mind, but as affording

[55] Studies on clinical populations might also support the claim that emotions are naturally tied to context. Rottenberg, Gross, and Gotlib (2005) found that individuals suffering from depression are much less sensitive to the context when attributing an emotion.

> the safest basis for generalization on the causes, or origin, of the various movements of expression. (Darwin, 1872/2009, p. 24)

The underlying assumption in this approach is, therefore, that animals do have emotions, and that these emotions share important similarities with our own. In Darwin's view, instead of supposing that human psychology was different from that of animals only in degree, but not in kind. Thus, we must look at the properties shared between our own psychological capacities, including emotions, and theirs.

At the folk-psychological level, attributing emotions to non-human animals is not a rare event. Wilkins, McCrae, and McBride (2015) investigated different factors that might explain why we attribute emotions to animals and which emotions do we attribute the most. The animals they included in their list were mammals (ranging from rats and squirrels to cats, dogs, and horses), birds (pigeons, chickens, and parrots), reptiles (cobras, crocodiles, and tortoises), fish, and invertebrates (cockroaches, fruit flies, and honey bees). They found that belief in the idea that animals have minds is the most important predictor of emotion attribution in animals in general. In other words, if we believe that animals have minds, then we tend to be willing to attribute certain degree of emotionality. So far, this is not very surprising. More interestingly though, they found that primary emotions (sadness, anger, joy, and fear) are attributed with significant frequency. Secondary emotions (guilt, pride, and jealousy), although less frequently, are also attributed in some cases.

Morris, Doe, and Godsell (2008) replicated similar findings. In this study, the investigators compared attributions between primary and secondary emotions to animals ranging from birds and rats to cats and dogs. They included not only the emotions mentioned above, but also primary emotions like anxiety and surprise, and secondary emotions such as grief, empathy, embarrassment, and shame. Their findings show that primary emotions are more frequently attributed than secondary emotions, with the exception of jealousy. Nevertheless, all emotions tested were attributed at least with some frequency and with relative confidence.

These studies show that we do in fact attribute emotions to animals. What drives these attributions? As I explained above, Wilkins, McCrae and McBride found that belief in animals having minds is an important predictor, although this may be easily explained by the fact that when we project our mental lives onto animals, we are open to the idea that they share many of our mental capacities. However, the fact that only primary emotions are attributed with a relevant degree of frequency and confidence shows that we do not hold animals capable of the same range of emotions as ourselves. What these studies show is that we project those emotions that we consider primary, that is, less cognitively demanding and less ingrained in our social lives. For instance, in Morris, Doe, and Godsell, social and moral emotions such as embarrassment and shame are the least frequent. In a similar vein, animals with less similarities to us such as birds and invertebrates are often considered less emotional than mammals.

Other factors that influence emotion attribution to animals include familiarity and ownership. Wilkins et al. showed that if participants are asked to rate emotions in animals intended for use (e.g. lab mice, pigs for meat) or pests (e.g. cockroaches,

rats), they attribute less emotionality overall in comparison to pets. The researchers explain this effect by appealing to the cognitive dissonance between our attribution of mental states and emotions, and our willingness to use and kill these animals. Whatever the explanation may be, this shows that the closer we feel to animals, the more psychological capacities we are willing to attribute. This, however, does not depend on the number of animals one owns or is related to. Morris, Lesley, and Knight (2012) showed that pet owners attribute a wide range of emotions to a number of animals including dogs, cats, and horses, but this effect is present if participants have at least one animal (there is no difference if they own more than two).

Lastly, it is interesting to see that the neural mechanisms behind these attributions are the same as those involved when we attribute emotions to humans. Spunt, Ellsworth, and Adolphs (2017) investigated these mechanisms using fMRI. They presented subjects with pictures of human faces, non-human primate faces, dogs, and a control scrambled picture. In one implicit task, they just presented the stimuli inside the scanner. In another explicit condition, they presented the stimuli and asked subjects to tell whether the stimulus matched with a description. In one condition, the description describes the emotion, asking whether the face is one of boredom, sadness, excitement, etc. In another condition, the description describes the expression itself, asking whether the human or animal is baring teeth, gazing up, opening its mouth, etc. They found that subjects did match the emotional description in all cases, and that the neural mechanisms involved in attributions to humans and animals were relevantly similar. These include activation in the dorsomedial prefrontal cortex (dmPFC), lateral orbitofrontal cortex (LOFC), and anterior superior temporal sulcus (aSTS). These areas are involved in theory of mind and reward processing.[56] Hence, the researchers conclude that there is no human-unique mechanism underlying emotional attribution, but rather a general mechanism that also applies to animals.

Besides purely folk-psychological attribution, it is worth mentioning that many of our scientific theories and discoveries regarding emotions are based precisely on these pretheoretical criteria of attribution as well. Consider studies such as those on fear conditioning in rats. Classical paradigms to study this phenomenon often use a stimulus that, prior to conditioning, evokes a particular response (i.e. an unconditioned stimulus). For example, when a rat is given an electric shock, it freezes and its heart rate and blood pressure elevate. Now, when researchers pair this shock with a tone and give the rat some time to learn this pairing, the rat starts freezing and presenting the same defensive behaviors even in the absence of the shock (i.e. the shock becomes a conditioned stimulus). This paradigm has been used to study the neural mechanisms underlying fear responses, and lead to the idea that the amygdala is involved in fear reactions (LeDoux, 2000).

This paradigm is based on the idea that the sort of response evoked by the shock in the rat is indeed a fear response. Why should we think so? On the surface, we call this a fear response because there are important similarities between what we call fear

[56] The researchers found some differences in activation between conditions, but they explain them as differences in conceptual familiarity, rather than in attribution.

in humans and the rat's behavior. For example, when we feel fear, we tremble, we feel our hearts pumping, etc. Furthermore, we also feel fear when shocked and other analogous situations. In this sense, we easily extend our concept of fear to the rat, and claim that we are evoking a fear response which we will later investigate. Similar to Darwin, we project our emotion concepts to animals in virtue of certain patterns that are shared across species.

Only after some reflection did researchers question the idea that the concept of «fear» applies to the rat in these cases. In this particular case, the reasons adduced are that fear necessarily involves a feeling to which we don't have access in the case of the rat, and thus we cannot justify this application (LeDoux, 2012). Nonetheless, this requires a particular theory of what emotions involve. Insofar as we are concerned with our folk-psychological attributions, it seems plausible that we can easily think (rightly or wrongly) that rats do have fear responses in virtue of its behavioral cues.

6.2 How do emotions fit together?

Having gone over the diagnostic features of emotions, now it is time to raise the question: how do emotions fit together? In other words, where does the unity of emotions lie? First, let us recap what I take to be the most prominent diagnostic features of emotions according to our folk-psychology. These are:

Arousal Emotions involve some state of arousal or intensity.

Valence Emotions are related to things seen as beneficial or harmful to ourselves. They inform us of our relations with objects in our environment.

Embodiment Emotions are associated with activity or feelings in different parts of our bodies.

Motivation Emotions motivate action. They lead us to act in particular ways, to react to different situations in a particular fashion.

Phenomenality Emotions thought of as involving feelings or experiential states.

Intentionality Emotions have intentional objects. They are about objects or events in the environment.

Flexibility Emotions are flexible in several respects. Emotions are triggered by a variety of objects, and expressed in a variety of ways. Additionally, emotion concepts are fuzzy, and they change across different cultures and languages.

Context-sensitivity Emotions are sensitive to the context in that one reaction may be categorized as an emotion or another depending on factors beyond our bodily reactions.

Attribution to non-human animals Non-human animals are described as having at least some emotions which are considered less cognitively demanding and that do not involve social norms.

With these features in mind, what is the best way to describe them in a scientific framework? Which type of scientific kind best accommodates to this phenomenon? In what follows, I will discuss each type of model according to the taxonomy of scientific kinds present in chapter 4. Since I intend to argue for a functionalist model, I will offer reasons to eliminate the other three models first. Before this, however, it is important to evaluate which features are the most important to tackle this question. In my view, not all features are central to our discussion. This is because either they are easily accommodated by all of the models under consideration, or because they do not offer a useful standpoint to abstract from them into scientific kinds.

6.2.1 Considering the relevance of folk-psychological features

As I explained above, some of the features I have listed can be accommodated by all models under consideration. In my view, this is clearly the case of arousal, motivation, and less straightforwardly, valence. Other features, as I will show below, can be set aside since they do not offer grounds on which to construct scientifically meaningful concepts of emotions. Specifically, I believe this is the case of phenomenality. Features that are accommodated by all models or that are irrelevant to the question at hand can be left out of the discussion, as they do not help us compare the advantages and disadvantages of different models of emotions. With this in mind, let us briefly go over each of these features and separate those that I will consider central to my argument from those that can be provisionally ignored.

First, let us consider features that can be accommodated by all models. In my view, these are arousal, motivation, and valence. Arousal can be understood in terms of patterns of physiological arousal, including increases in heart rate, skin conductance, and the like. Essentialist and historical models can include these reactions as part of the essential or evolved mechanisms corresponding to an emotion, respectively. In the case of functional models, they can either accommodate arousal also in physiological terms as the realization of a functional profile, or, as social models can also do, in terms of dispositions to engage in energetic behavior or practices. In any of these cases, arousal seems to pose almost no problem.

Motivation can also be accommodated in a number of ways. We can understand emotions motivating action in terms of the effects of neural and physiological patterns that would presumably individuate emotions in the case of essentialist and historical patterns. This is how basic emotion theorists, particularly those appealing to affect programs, cash out emotional motivation. According to these theories, affect programs are instantiated in circuits which lead to automatic forms of action, explaining the motivational role that emotions play. These circuits may be then understood as part of an emotion's essence or as part of the realizers that have evolved to support it.

When it comes to functionalist and social frameworks, we can spell out the fact that emotions lead to action in terms of the behavioral dispositions involved in the functional description of an emotion or in the social practices that would identify each emotion, respectively. For the functionalist, the emotion itself would have behavioral dispositions as constituents. For the defender of a social model, emotions could also

be individuated in terms of behavioral dispositions, only that this time they are dispositions to engage in a determinate set of social practices.

When it comes to valence, things might not be so straightforward. Yet, it is plausible to offer a story about how valence fits into these frameworks. On essentialist construals, an emotion's valence must be cashed out in terms of how an emotion's mechanisms produce states of pleasure or displeasure, as well as avoidance or approach reactions. If an emotion's mechanisms lead to displeasure or avoidance, this suggests that the emotion is a negative one. Conversely, if it leads to pleasure or approach, we can think of the emotion as positive. Even though such an account requires specifying how pleasure or displeasure can be mapped onto specific and consistent structures in the brain and body, it is presumably possible to offer such an account.

Regarding the remaining frameworks, similar stories can be offered. Historical models are well prepared to accommodate valence, given that they can appeal to a historical background that has led to an emotion having adaptive value. If the emotion's adaptive value is tied to the avoidance of objects in the environment, we can say the emotion is negative; in the opposite case, if an emotion is adaptive because it encourages approach behavior, the emotion can be said to be positive. Functionalist models can also appeal to dispositions of approach or avoidance behavior to draw a similar distinction. Lastly, social models can appeal to social values, hence accommodating valence without insurmountable problems.

Besides these features, it is also worth considering features that may not be informative even if explicated. As I explained above, I believe this to be the case of phenomenality. As I argued in section §2.2.5, descriptions of an emotion's phenomenology can lead to two outcomes. First, if we interpret phenomenality in terms of qualitative, subjective experience, then any explication of phenomenality must offer an answer to the explanatory gap problem. While this could be in principle possible, as I claimed before, this is still contentious.

Second, a different approach would be to understand an emotion's phenomenality in terms of second-order states. As I argued in the discussion on phenomenological patterns, these descriptions end up collapsing into behavioral dispositions and even physiological outcomes at best, features that we have included in terms of motivation and embodiment, or again end up demanding a solution to the explanatory gap problem altogether. Thus, on this second interpretation, phenomenality is not an independent feature in the best case, but rather points to forms in which other features figure in how we describe our emotional episodes in our folk-psychology.

If these arguments are correct, then the most relevant features to decide between different models of scientific kinds are embodiment, intentionality, flexibility, and context-sensitivity, as well as how emotions are attributed to non-human animals. In the next section, I will then compare each of the frameworks with regards to how well they accommodate these features. In other words, I will examine which type of model provides the best candidate to explicate emotion categories into scientifically tractable kinds.

6.2.2 What emotions are not

Essentialist models

Traditional theories emotions have assumed that the unity of emotions lies in a hidden essence, perhaps in specific and consistent patterns of neural and physiological activation (Barrett, 2006; Panksepp, 2008). This way of understanding emotions construes them then as essentialist kinds. The idea is that an instance of an emotion a member of that emotion category if it is produced by the same underlying neural or physiological mechanisms as other members in the category. In the most general case, an essentialist model of emotions appeals to a common underlying concrete structure to all instances of an emotion category.

In my view, critics of traditional accounts of emotions are right in their skepticism about this construal. On one hand, empirical evidence on the realizers of emotions does not support the view that emotions can be individuated according to underlying essences. As the debate stands, this seems to be at least an approximate consensus. Most of the evidence for concrete structures underlying emotional reactions suggests that mapping emotions onto specific and consistent brain regions or physiological measures have failed, as I showed in chapter 2.

Furthermore, however, I believe that our folk-psychological understanding of emotions does not support the essentialist model either. The way in which we categorize emotions in our folk-psychology is not due to hidden essences in terms of neural or physiological mechanisms, or other mechanisms for that matter. An essentialist construal of emotions would have problems explaining emotions' flexibility and context-sensitivity. According to such a construal, the flexibility of emotions would have to be described either as a divergence from standard cases that do possess the essential features of emotions, or as different emotions altogether.

Consider two instances that we would classify as fear: fear of snakes and fear of failing an exam. Fear of snakes, which is the canonical example of essentialist models of emotions, can presumably be cashed out under the essentalist scheme by appealing perhaps to amygdala activity. Let us assume this is correct. What about fear of failing an exam? We can easily imagine someone being afraid of failing an exam without constant amygdala activity.

On an essentalist account, such a case would have to be construed either as a deviation from the standard, or as a different emotion altogether. On the first horn, most of our everyday instances of emotions would count as non-standard, given the variety of ways in which emotions obtain. Any non-basic instance of fear would be a non-standard instance. This is undesirable as it would be unclear how the standard is fixed in the first place, since most of the actual explananda phenomena (our everyday fear behavior) would be considered non-standard. On the second horn, we would be forced to split the two emotions even though there is a clear sense in which we describe the latter case one of fear. If any slight variation leads to systematically splitting the kind, then we risk exploding our taxonomy of emotions and hence creating an problematic mismatch between our folk and our scientific taxonomy.

To make this case clearer, consider an analogy with a quintessential essentialist kind: H_2O. Instances of water that we encounter in our everyday interactions, and which fix our pretheoretical concept of WATER are not substances containing only H_2O. In a sense, what we call "water" in our everyday interactions, $water_P$[57], as it were, is not the same as the water referred to when we use the concept in a scientific way, $water_S$. This mismatch is clarified in ordinary language by means of expressions such as "Tap water is not *pure* water." For most cases, this clarification works fine: tap water is not a standard case of the kind we make reference too when we refer to water in a scientific sense.

If we apply this to emotions, however, the problem quickly comes to the surface. For many cases in our everyday interactions, we will be forced to say that they are not "pure" instances of fear, anger, sadness, etc., and only those conforming to the essentialist construal will qualify as "pure." This way of understanding emotions renders most of our pretheoretical concepts, which I have argued fix our explanandum, "impure" cases. A scientific theory of emotions that renders most of our emotions impure, I submit, is not a good theory. What we want is a theory that takes emotions in our everyday interactions seriously, since that is precisely what we want to explain.

Taking the other strategy to deal with this, dividing emotions into different scientific concepts of emotions, is also problematic. This would quickly lead to a change of subject, since most of our ordinary emotion concepts would have to be divided and lumped together in a taxonomy that no longer would be translatable to our folk taxonomy. Hence, there would be good reason to doubt that our theory refers to emotions as an explanandum at all. In sum, in any of these cases, cashing out emotions in terms of hidden essences yields an overly restrictive framework that does not fare well when contrasted to what we have identified as the explanandum of our theory.

One possible reply is that while empirical evidence does not support past essentialist construals, it is not in principle impossible that we will be able to offer one. In other words, we must only postpone the identification of essential properties tying emotions together. In my view, this strategy is misguided for various reasons. First, as I will defend below, there are other models that currently fare better with empirical evidence and that preserve our pretheoretical taxonomies better than essentialist construals. In terms of comparison, we have better alternatives on the market.

Second, as I discussed in chapter 4, there are problems with essentialism about kinds in general, as it seems that essentialism does not capture actual scientific practices and metaphysical assumptions (Khalidi, 2013). Furthermore, similar to the comparison regarding empirical evidence, we have other accounts of kindhood available that are better suited to explain how scientific taxonomies work. Here I have defended a pluralistic account of kinds which admits other models. Given these alternatives, remaining tied to essentialism is not warranted.

Lastly, and most importantly, essentialist models would force ignoring empirical evidence until we find evidence in its favor, when we already have evidence that emo-

[57] I will use the subscript P to distinguish pretheoretical concepts from scientific ones. I will use the subscript S for the latter.

tions do not correspond to specific and consistent microstructures that would qualify as the essences of emotions. Keeping an essentialist model would deny evidence in favor of variability, which I showed in chapter 2 to be convincing. The essentialist would not only be postponing an account of emotions until evidence in their favor is found, but would be denying that past evidence speaks against their view.

At best, essentialists could relax their criteria for what counts as an essence. I have taken essences to be necessary microstructural properties. In the case of emotions, I have linked them to specific patterns of neural or physiological activity. The essentialist could reply that we need not map emotions onto essences in these terms. Perhaps an emotion's essence lies in a corresponding functional network, as found by multivariate analyses. This would imply relaxing the concept of an essence to admit higher level properties.

While this move may sound convincing, I believe there are also reasons to resist it. Relaxing the criteria for essences to admit higher level properties would risk rendering essences uninformative, as essences would either collapse into other types of kinds or would apply to any arbitrary definition we can offer. For any set of objects, we can define a higher level property that we could count as their essence, hence counting this set as an essentialist kind.

Historical models

Crossing out essentialist kinds as a candidate to construct emotions as scientific kinds, we are left with historical, functional, and social kind models. Each of these models have noticeable advantages when compared to the essentialist model. All of these models can deal better with variability. A historical construal would accept variability as long as causal history remains constant, i.e., instances of emotions with a common phylogenetic and/or ontogenetic developmental history would still qualify as members of the same category in spite of variations. This would work akin to biological concepts for different species. Two instances of the class TIGER would be considered tigers in spite of them differing in important respects as long as there is an evolutionary line connecting the two. Regarding functional kinds, their compatibility with multiple realization buys us variability in terms of different forms of realization of a given emotion category. As long as there is a common functional pattern that can describe commonalities across instances, variations are not only tractable, but expected. Lastly, social kinds would admit variation in terms of variations in our social world. Since our social institutions are fluid, malleable, and changeable, it is also expected that emotions, construed as such, would exhibit different forms across instances. Given these advantages, let us explore each type of model in turn.

Historical models of emotions, besides allowing some degrees of variability, also enable us to fit emotions into an evolutionary framework. This makes them attractive for those interested in the role of emotions in our ancestry. For example, a historical construal of emotion would serve to motivate hypotheses regarding the origins of language and communication (Bar-On, 2017), or to ground theories of emotions along the lines of Darwin's views.

One such historical model of emotions is the one suggested by Griffiths (1994, 1997). Griffiths suggests that emotions—specifically those captured by Affect Program Theory—are better cashed out using a cladistic approach. He argues that the cladistic approach can go beyond merely describing emotions in terms of tasks, since it enables questions about underlying causal mechanisms. This is because, in contrast to alternative approaches such as an ecological or a functionalist approach, a cladistic approach can use the resources of evolutionary theory to explain what emotions are and how they have come to be. In his words:

> We do not yet have a general theory of the relationship between organism and environment ("general" meaning that it applies to diverse lineages of organisms). The historical nature of the evolutionary process may mean that no such theory is possible. We do, however, have a sophisticated practice for determining evolutionary relationships (homologies). We know that this system of classification reflects the actual process which accounts for the diversity of life and the clustering of individuals into groups with shared features. This is as true of features at the level of task description as of features at the computational or implementation levels. To attempt to develop a science of emotion using general ecological categories instead of categories of homology would be to choose a set of categories with a poorly understood and possibly very weak causal homeostatic mechanism over a set of categories whose causal homeostatic mechanism approaches the traditional ideal of a natural kind. (Griffiths, 1997, p. 240)

I find Griffiths's approach inadequate for a number of reasons. As I already explained above (see chapter 3.2), Griffiths fails to address why a general theory of emotions is impossible in principle. The argument offered here has the same flaw. The fact that we do not yet have a non-cladistic (i.e., ecological or functionalist) theory of emotions does not entail that such a theory cannot be offered. While this may be true for other accounts as well, there are additional reasons to resist historical construals such as Griffiths's.

First, Griffiths's view is explicitly premised on the idea that a proper account of emotions must cash them out as kinds in terms of Boyd's HPC account. In chapter 4, I offered reasons to think that this is an overly restrictive approach to kinds. If we abandon the idea that all scientific kinds must be understood in the same framework, Griffiths's commitment to Boyd's account becomes problematic. While it might be true that if emotions were to be understood on such an account, the best approach would be a cladistic one, we can resist the premise that we must spell out emotions as HPC kinds. With more kinds to choose from, there is no need to follow Griffiths in this regard.

A consequence of shifting to a pluralistic account of kinds brings out a second reason to resist Griffiths's arguments. Griffiths's skepticism about the ecological and functionalist projects is overly speculative. He limits himself to claim that we do not understand ecological and functionalist categories as well as homological ones. The

only reason to prefer homological categories is that they fit better with the HPC account. Again, I resist such a commitment. Moreover, I find the functionalist project more promising than what Griffiths does. In my view, there are a number of successful generalizations on functional kinds. These even include categories in physics such as "machine" and psychiatric categories such as "depression." Hence, I do not find Griffiths's pessimism about functional kinds warranted. If there are genuine scientific kinds that generalize on functional profiles, then they must still be considered as candidates to construct emotion kinds.

So far, my arguments only cast doubt on Griffiths's recommendation to use a cladistic approach to construct emotions as historical kinds. Now let me offer reasons to resist the historical model altogether.

First of all, I believe historical models still have problems when dealing with some forms of variability. Even though they can handle some degree of variability relative to essentalist models, they are still limited in this regard. Consider two instances of an emotion we are tempted to label 'fear.' Let us assume that these two instances have different underlying mechanisms with different phylogenetic histories. On a historical model of emotions, this would force us to categorize these instances into two different kinds. Yet, we can raise the question of what makes it that we are tempted to label them instances of 'fear.'

In my view, we can offer good reasons to categorize them as such. For example, if the subject in both of these instances displayed similar behavioral outcomes, it would make sense to say in both cases that they are afraid. We can imagine them trembling, sweating, and running away from danger, even if the mechanisms underlying these reactions differed in their phylogenetic origin. This, I believe, is how our folk-psychological attribution of emotions work. We do not attribute emotions to others by looking at the mechanisms underlying their reactions and from there apply one or another emotion concept. Instead, we look at how they behave in certain situations and attribute the corresponding emotion.

If this is true, then it follows that cashing out emotion categories in terms of historical pattern misses out on forms of variation that our folk-psychological attribution is well prepared to handle. The fact that two instances of an emotion vary in their underlying mechanisms and historical origins is not a reason to split categories. This fact about our taxonomies of emotions cannot be accommodated under a historical model, since the historical model forces us to divide up kinds when we normally would not.

Besides this issue with variability, I believe historical models also have problems when it comes to explaining social factors involved in our folk emotionality. For example, explaining shame and embarrassment in historical terms would demand a historical account of social and moral norms in order to cash out these emotions as reactions to perceived failures to meet them. Even if such an account could be offered (see Schlingloff & Moore, 2017, for reasons against this), this would hardly mirror ordinary instances of these emotions. For instance, consider an account of shame that would apply across our evolutionary history. Even if this account could fit shame into our

overall story regarding our ancestry, it would not easily account for more prototypical cases of shame such as feeling ashamed that we have stood up a good friend or that we have not called someone that was expecting us to do so.

Overall, historical models do not fare significantly better than essentialist models. Even though they admit some degree of variation, they do not admit the kinds of variation present in our folk-psychology. Additionally, they raise difficulties when it comes to explaining some of the more socially complex emotions by demanding a historical account of social factors. Since these cases constitute part of our explanandum, historical accounts, while more permissive than essentialist ones, still come out as too restrictive for the purposes of our theory.

Social models

The arguments above leave us with two contenders on the ring: functional kinds and social kinds. As I have already signposted at several points in this work, I believe that functional kinds are a better framework than social kinds. To argue for this claim, let me first consider the advantages of social kinds over the other models, so as to motivate the claim that functional kinds also share many of these advantages plus others.

First, an account of emotions in terms of social kinds would be well prepared to handle high degrees of variability. Such an account would find unity in terms of underlying mechanisms fairly irrelevant, since unity would obtain at the level of social practices. Whether or not a given emotion has a homogeneous set of underlying mechanisms is not an important question (even if we found such homogeneity). Furthermore, such a model would be well suited to handle the social aspects of some emotions as in the cases of shame and embarrassment presented above.

Other advantages of a social model of emotions would be that it can easily accommodate the flexibility of emotions, their valence, and their context-sensitivity. If emotions were understood as kinds of social practices, it would be easy to explain why they are triggered by different objects, why they are expressed in different ways, in which sense they can be positive or negative, and why they vary depending on the context. For example, we could offer an account in which emotions are learned reactions to objects that are socially positive or negative, hence explaining their variability in terms of objects and explaining valence. Furthermore, we could invoke rules of expression and behavior to account for how they are expressed in different contexts as a function of the social norms that are at play. Additionally, a social model of emotions could also easily explain why and how our emotion concepts vary throughout cultures by appealing to the different practices involved.

In chapter 4, I presented Khalidi's distinction between different types of social kinds. These are:

- Kinds that require mental attitudes to be in place, but not necessarily towards the type nor towards tokens (e.g. racism).
- Kinds that require mental attitudes to be in place towards the type, but not towards the token (e.g. war).

- Kinds that require mental attitudes to be in place towards the type and the tokens (e.g. permanent resident).

When evaluating social models of emotion kinds, we can raise the question of which type of social kinds would suit emotions best. First of all, in any of these cases, the claim that emotions form social kinds amounts to the claim that they require mental attitudes of some sort to be in place. This is because these mental attitudes will provide the basis to constitute emotions as social practices. A social model of emotion kinds could cash out these attitudes in terms of evaluative attitudes. What emotions require, in this sense, would be mental attitudes towards the objects they are about. This would integrate intentionality into the account from the outset.

Nevertheless, this already introduces a problem, namely, that is makes emotions highly cognitively demanding. If emotions require evaluations to be understood as attitudes, then emotions require the cognitive capabilities that attitudes require, arguably including propositional thought and shared concepts. While this might be true for some emotions anyway (especially social and moral emotions), it is not a plausible claim for basic cases such as fear of snakes or anger bouts. These emotions are best understood in terms of automatic mechanisms, which is why they are the gold standard for basic emotion theories. Spelling them out in terms of mental attitudes makes them more cognitively complex than is required to explain how they work. Let us however assume that this can be addressed and further discuss the plausibility of such models.

Provisionally conceding that all emotions could require mental attitudes to be in place, the next question is whether they require attitudes towards types, that is, whether emotions require attitudes towards emotion categories. Recall the example of war presented in chapter 4. For there to be instances of war, there must be attitudes towards battles and conflicts which we categorize as wars. It is possible however that there are wars that, at the time they obtain, are not categorized as such. Hence, for there to be wars there must be attitudes towards the type but not towards specific tokens.

Assuming that emotions require mental attitudes towards the type does not seem to demanding at first. Social accounts of emotions might stress that this precisely shows why emotions very across cultures, since different societies might have different attitudes and hence different taxonomies. Yet, this requirement would entail that before emotion concepts were developed, there were no emotions in the first place. Attributing emotions to societies without such concepts would be anachronistic. This becomes particularly problematic if we want to integrate emotions into an evolutionary story. If emotions are to have adaptive value in the sense of providing automatic evaluations of the environment, a social account of this sort becomes highly implausible. The reason for this is that in this case, early groups that lacked emotion concepts would not have the right type of attitudes required to have emotions in general, hence precluding an evolutionary explanation of emotions.

This problem becomes even more dramatic if we claim that emotions require mental attitudes both towards types and tokens. In this case, we would require not only that

there are types of which we are conscious, but that each time someone felt emotion, there would be someone attributing that emotion to them. Consider the analogy with the example of this type of social kind above, permanent resident. Without someone in an institutional framework declaring a person a permanent resident, a person cannot be said to be a permanent resident at all. If this were the case for emotions, emotions would necessarily require others to attribute the emotion for an emotional episode to obtain.

Someone might reply that such attribution can be self-referential. In other words, it is possible that an emotion obtains if one attributes that emotion to oneself, thus satisfying the criterion that there is some attitude towards the token. However, this would imply that we cannot be wrong about our own emotions. If I can successfully instantiate any emotion that I attribute to myself, there is no possibility of error. This claim is highly questionable, as there are a myriad of cases both in our everyday lives and in controlled settings in which we are wrong about our own emotions. This is precisely why we seek therapy to recognize and manage emotions that we are unclear about. Hence, at the very least, a defender of this sort of account must either accept that we are infallible regarding our own emotions, reject that attribution leading to the instantiation of an emotion can be self-referential, or offer an account of error that is compatible with this sort of self-referential attribution. As I have claimed, the first seems improbable. Regarding the second, this would require emotions to always be instantiated in the presence of others, which is quite counter-intuitive. Finally, concerning the third option, I do not see how such an account can be offered.

In sum, social models of emotions can make recourse to two types of social kinds: kinds requiring attitudes but not towards types or tokens, or kinds requiring attitudes towards types but not tokens. Both options require a degree of cognitive sophistication that might run counter to an evolutionary picture of emotions that claims that emotions have adaptive value. It would force either giving up the integration of a theory of emotions with an evolutionary story, or explaining which and how the mental attitudes required for emotions arose and conferred an evolutionary advantage while still constituting social kinds.

Additionally, this also makes it difficult to account for the folk-psychological attribution of emotions to non-human animals. In my view, we should consider this problem seriously. The main reason for this is that, as I showed above, attributing emotions to non-human animals is part of how we attribute emotions in our daily lives, and hence qualifies as part of the explanandum of our theories of emotions. A satisfactory theory of emotions, in this sense, should at least be compatible with the possibility that animals do have emotions, even if it turns out as a matter of empirical fact that they do not. A theory that from the outset posits overly demanding constraints that preclude animals from having emotions would not reflect an important part of our folk-psychological understanding of the phenomenon.

Besides this degree of sophistication, social models must clarify whether the mental attitudes that need to be in place for emotions to obtain are also attitudes towards types. While an account in this direction might still be possible, I believe there are

good reasons to doubt of its promise. Again, this makes it even more difficult to integrate emotions into an evolutionary picture and runs counter the claim that subjects may have some emotions (e.g. fear) even if their social background lacks the require concepts.

Lastly, I believe that social models of emotions overstate variability in a number of ways which may turn out to be problematic for a number of reasons. Social accounts of emotions, such as the one offered by Lutz (1988), claim that emotions are social phenomena because they involve different objects and concepts in different cultures. One of the examples driving her view is the case of *song*, which she translates to "justifiable anger." The following anecdote illustrates this emotion:

> While in the field, I had twice been mailed a bottle of liquor by well-meaning friends. The first—a bottle of wine—I shared with the two women who had become my closest friends, but greatly regretted doing so the next day. As women do not drink except in secret and when given the local coconut toddy by one of the men who manufacture it daily, and as we had the misfortune to be observed, Tamalekar and the brothers of the two women were justifiably angry (*song*). When a bottle of whiskey arrived for me many months later, I decided that the best course of action would be to give the bottle to Tamalekar. As he drank from it that evening and became progressively more intoxicated, he repeatedly and it seems pointedly made reference to the fact that his daughter had given him a bottle of liquor. (Lutz, 1988, p. 37)

In Lutz's view, *song* is different from anger in that it is bound to moral transgressions, not to events that people merely dislike (as in anger). Because of these differences, Lutz claims that we can only understand *song* in the cultural context in which it obtains. In her book, Lutz presents other interesting examples to support her claim.

I believe Lutz points out interesting facts about emotion concepts and their cultural variations. A scientific account of emotions should take into consideration the fact that an emotion can be elicited and described in different terms and with different concepts across cultures. Yet, I do not think this implies that emotions can be reduced to social practices.

The examples which Lutz discusses are examples of emotions with two variations relative to a Western characterization: objects and expressions. In the case of *song* and anger, what differs in the first place is the object of the emotional reaction. Whereas in the case of *song*, objects of the emotion must involve moral transgressions, objects of anger do not. Yet, both emotions share important features, albeit more abstract ones. Both are negative emotions, and both are elicited by a sense of transgression and offense. Thus, there are descriptions that capture both emotions while admitting that they vary in their objects. Moreover, these descriptions are what enables Lutz to compare the two emotions. Without such a description, no translation between the two would be possible.

A second type of variation that Lutz discusses extensively in her work are variations in expression. Following the case of *song*, Lutz notes that anger in Western contexts is an emotion whose expression ought to be suppressed. As she explains, expressing anger—at least violently—is a sign of immaturity. In the ideal case, people are expected to suppress their anger and calmly resolve conflicts. This is not the case among the Ifaluk, the community that Lutz studied. While the Ifaluk are a pacific society, they expect others to express *song*, especially adults. Expressing *song* is taken as a sign of maturity, of sensibility to moral transgressions.

Taking variations in expression of this sort as criteria to divide the two emotions overstates the differences and neglects important similarities. In both cases, there are common ways of expression which Lutz recognizes. These are:

> [...] refusal to speak or, more dramatically, eat with the offending party; dropping the markers of polite and "calm" speech; running away from the household or refusing to eat at all; facial expressions associated with disapproval, including pouting or a "locked" mouth, "lit-up" or "lantern" eyes; gestures, particularly brusque movements; declarations of song and the reasons for it to one's kin and neighbors; throwing or hitting material objects; and in some cases, a fast or the threat of suicide or other personal harm. (Lutz, 1988, p. 174)

All of these reactions are also present in anger. What is different in the two cases is what is considered a socially adequate response. For Western societies, any openly emotional expression is to be suppressed, and a "cold-headed" form of expression is to be expected. For the Ifaluk, this is not the case. Yet, there is a relevant sense in which we can say that both societies experience the same emotion, but that they have different norms regarding how it is to be handled.

The objections against Lutz's view can be generalized to other purely social models of emotions. By focusing only on the social variations of emotional behavior, social models neglect similarities in terms of broad patterns of behavior. This leads them to a problem of translation. If there are no similarities between Western concepts of emotions and other cultures' concepts, then there is no sense in which we can understand their behavior as emotional. Conversely, if we can understand their behavior as emotional, it must be because there is something relevantly similar between instances of their emotions and ours, in spite of cultural differences. Hence, I think, we must look for a description that allows us to admit variations while making these similarities explicit.

It is true that social models can be amended in order to avoid some of these shortcomings. We could construct a social model of emotions that does not overstate variability, yet generalize across emotion categories individuated in terms of social practices. Nevertheless, I believe that social models are still at a disadvantage when compared to functional models. While emotions do involve social components that are important to consider, we need not individuate them because of their sociality.

In my view, strictly speaking, we can apply emotion concepts independently of social practices. Consider a person in isolation who sees an dangerous animal approaching or that has run out of food and sees their survival threatened. Regardless of their social context, which in this case is absent, there is a clear sense in which we can say this person is afraid of the animal or anxious about their means of nourishment. We can do this because, plausibly so, this person would have dispositions to behave in ways which we can describe as fearful or anxious. For example, this person would run away from the dangerous animal, or urgently seek for food by any means available. These dispositions warrant the application of emotion concepts, suggesting that we can individuate emotions without social practices involved.

Someone might object, however, that there are emotions which cannot be described without appeal to the social context. These may be cases of embarrassment, for instance, understood as a hiding reaction due to the belief that one has failed a social norm. Without recourse to social norms, the objection would hold, we cannot make sense of this emotion. While I concede that this may be the case for social emotions, I believe these factors can be integrated into a functional framework which applies as well to the simple cases I appealed to before. This is because functional descriptions can also accommodate social practices while not being committed to the claim that social practices are necessary in every instance of an emotion. If this is true, this shows why functional models are more powerful than social models, since they allow generalizations that social models would preclude. To see this in detail, let us elaborate on how a functional model can be cashed out.

6.2.3 Why a functional model?

At last, we come to functional models of emotions. As I understand functional kinds, they are kinds projectible in virtue of a common functional profile. We can characterize a functional profile in terms of relations between inputs, intermediary states, and outputs (Block, 1978; Putnam, 1960/1997). These outputs are taken to be the contribution of the functional system to the overall behavior of a system, as I discussed in chapter 4. Examples of these kinds include the aforementioned cases of machines in physics and psychiatric categories. A lever, for instance, can be spelled out as a physical system which takes in a force and outputs a different force. Different objects may be said to realize this function if, when engaged in a relevant system, the object transforms force in the way specified by the functional profile description.

As I explained before, functional kinds allow multiple realizability, in the sense that it is possible for a number of realizers to fulfill the functional profile. For example, in the case of the lever, levers can be made out of a variety of materials. What makes an object be a lever is that it allows force to be transferred in a particular way, independently of what the object is made of. This is why functional descriptions can be understood, following Roth and Cummins (2017), causal relevance filters, since they filter out details about the realizers which are irrelevant to the object carrying out its function (e.g. a lever's color).

One way of cashing out emotions as functional kinds, following Prinz (2004), is by individuating them in terms of Lazarus's *core-relational themes*. On this view, an emotion is, first, a relation between an organism and its environment. Different emotions involve different such relations. Lazarus (and Prinz) present the following core relational themes:

Anger A demeaning offense against me and mine.

Anxiety Facing uncertain, existential threat.

Fright Facing an immediate, concrete, and overwhelming physical danger.

Guilt Having transgressed a moral imperative.

Shame Having failed to live up to an ego-ideal.

Sadness Having experienced an irrevocable loss.

Envy Wanting what someone else has.

Jealousy Resenting a third party for loss or threat to another's affection.

Disgust Taking in or being too close to an indigestible object or idea (metaphorically speaking).

Happiness Making reasonable progress toward the realization of a goal.

Pride Enhancement of one's ego-identity by taking credit for a valued object or achievement, either our own or that of someone or group with whom we identify.

Relief A distressing goal-incongruent condition that has changed for the better or gone away.

Hope Fearing the worst but yearning for better.

Love Desiring or participating in affection, usually but not necessarily reciprocated.

Compassion Being moved by another's suffering and wanting to help. (Lazarus, 1991, p. 122)

These characterizations of each emotion can be interpreted as a part of the functional profile that individuates each emotion. In the first place, they point out a set of inputs for each emotion category. Anger, for example, requires the perception of an offense; envy, the desire to possess an object someone else has, etc. Yet, these descriptions are not sufficient to specify a full functional profile. In many of the cases above, what is described is merely a situation. Emotions are not just these situations. They are *reactions* to such situations. Anger, for example, is not merely the presence of an offense against me. It is a reaction to an offense against me. These reactions constitute the outputs of each emotion.

What characterizes these reactions? In my view, these reactions are, primarily, dispositions to engage in specific patterns of behavior. It is when we are bound act in particular ways to these situations that we describe ourselves as experiencing an emotion. To continue with the example of anger, not every reaction to an offense is an instance of anger. If I perceive an offense against me, but I am indifferent to it, we would hardly describe me as angry. Instead, what warrants describing me as angry is that I behave in a way that involves some form aggression or violence.

Consider a different case such as fear (or fright, in the list above). Describing someone as afraid means describing their state as being disposed to react to facing danger. The behaviors this person would be disposed to engage in might involve fleeing from the object of danger, fighting for self-preservation, or freezing and trembling. Again, other ways of dispositions would not constitute instances of fear, such as being indifferent or, to draw on more border cases, laughing in the face of danger.

Hence, we have two items in my account of emotions: dispositions to engage in specific patterns of behavioral reactions as outputs, and situations we react to that instantiate relations between the subject and its environment as inputs. There is yet another ingredient to be added to the mix. This is the fact that emotions involve objects, reasons, and also often involve beliefs and other cognitive states (though not always; see Pérez 2013). It is easy to imagine such cases. Consider feeling afraid of failing an exam, or being happy that one has succeeded in passing it. In both of these cases, there must be mediating states of believing that the exam was difficult, the desire to pass the exam, and the belief that one has passed the exam successfully in the last case. These can be introduced as intermediary states that allow the system to output the specific dispositional responses for each emotion category.

In the more complex cases, this feature of emotions has led some to believe that emotions are judgments (see e.g. Nussbaum, 2001; Solomon, 2003). In my view, these need not entail such a conclusion. First, not all emotions require propositional attitudes. To draw on an example by Pérez and Gomila (2014), someone can be afraid of spiders without having propositional attitudes about spiders. This would constitute a basic case of fear, but a case of fear nonetheless. Second, even though more complex emotions might require cognitive states, this does not mean that these emotions are merely judgments. Cases of shame or embarrassment arguably do require some mediating states such as the belief that one has failed to meet a moral or social norm, respectively. Yet, we can cash out these emotions as dispositions to react in particular ways such as hiding oneself under the belief that one has failed a moral or a social norm in a given situation.

Overall, what I take to be the moral of these observations is that a proper model of emotions must be compatible with their relations to other cognitive states to account for cases where this relation obtains, without making such relation necessary. In other words, a theory of emotions must spell out emotions in such a way as to allow emotions to relate to complex cognitive states, but not in a way that makes these states necessary. This would allow our theory to capture a variety of cases, ranging from

simple cases such as fearing spiders, to more complex cases such as feeling ashamed or embarrassed.

I believe that functionalism allows just that. On a functional description of emotions, we can admit that emotional states involve other cognitive states in some cases, but do not require them at the core. All that is necessary for an emotion to obtain is that there is some pattern of behavior in a situation that instantiates a relation between the organism and the environment that has bearing on its well-being.[58]

Hence, I submit that emotions can be spelled out as dispositions to engage in patterns of behavior which may involve other mental[59] states, and that are elicited by situations that involve specific relations between the organism and its environment. These elements—dispositions to engage in patterns of behavior, other mental states, and relations between the organism and its environment—are what constitutes a functional description of an emotion. They are an emotion's outputs, intermediary states, and inputs, respectively.

Cashing out emotions as functional kinds has clear advantages. First, as I said above, functional descriptions allow for multiple realization. This can help us accommodate the embodiment of emotions. In my view, emotions are embodied in that the body is the realizer of the functional profiles that specify each emotional state. For example, when we say that fear is characterized by trembling, it is a body that trembles. When we say that we have a tendency to attack when we are angry, it is our bodies that prepare to attack. What multiple realizability buys us is that we need not map emotions one-to-one onto specific bodily states. Emotions can be realized by a variety of bodily states, as long as these bodily states can be described as part of the satisfaction of the functional description at hand.

This way of taking advantage of multiple realizability already makes functional models more adequate than essentialist or historical models. The difficulties I raised against these latter models are not present for functionalist ones. Consider the case of fear in cases where one fears snakes or fears taking an exam. As I presented the case, the first might involve amygdala activity, while the other does not. On the essentialist model, categorizing both cases as fear demanded either fixing the first as the standard and the other as a deviant, or splitting the kind. On a functional model, this can be easily accommodated in terms of different instantiations of the same overarching patterns. What makes both cases instances of fear is that there is a perceived danger and a reaction to it (fleeing from the snake, sweating during the exam). Whether or not there is an common concrete underlying structure is not important, since it is the abstract structure that unites the two cases.

With regards to historical models, I argued that they ran into difficulties integrating social aspects of our emotional lives. Functional models, however, enable us to

[58] This line of argument separates my account from a purely behaviorist account. I am not claiming that emotions are just dispositions to engage in patterns of behavior, but I am allowing mediating states to be add in what characterizes an emotion.

[59] I do not restrict the mental states involved to cognitive states. Without an account of what constitutes a cognitive state (in contrast to other possible types of mental states), I will remain agnostic as to such a specification. To be safe, I shall use the term "mental" broadly.

take these aspects into consideration. In the discussion above, I claimed that for some emotions such as shame or embarrassment, historical models demanded a historical account of social and moral norms. We can see how functional models can accommodate these norms without such an account by noticing that functional descriptions allow for complexity and relations with other mental states. We can account for shame or embarrassment by describing them as emotions that require the belief that one has failed to meet a norm. How we acquire these beliefs is a task for social and cognitive psychology, but on this account, we need not reduce these to evolved mechanisms. A higher order account of the observation of norms would suffice to construct a model of these emotions at an abstract level. Hence, adopting a functional framework opens the door for models that overcome this disadvantage of historical ones.

Lastly, we can also see some advantages in contrast to social models. Functional models can resist overstating differences in the intentionality of emotions, which I argued was a problem for social accounts. As I explained above, social accounts of emotions make a jump from the fact that emotions differ in their intentional objects across cultures, to the claim that emotions differ categorically across cultures. A functional framework does not make this jump, since we can understand two emotions in different cultural contexts as functionally similar. In the case above, *song* and anger can be said to be the same emotion insofar as they are both negative reactions to an offense, be it a moral transgression or an event that frustrated our personal endeavors. On this abstract way of describing the emotion, both emotions share important features, such as violent reactions and attack tendencies. The fact that they are triggered by different objects in different cultures does not entail a split of the kind.[60]

Second, adopting a functional framework can help us make sense of why and how we attribute emotions to non-human animals, which I claimed was part of our folk-psychological attribution of emotions. On this view, when we claim that animals have emotions, it is because we notice that their behavior satisfies the functional characterization of a given emotion. For example, when we say that our dog is happy to see us or that our cat is angry, we do so because our dog exhibits energetic behavior or because our cat reacts violently and attacks, respectively. On these descriptions, we even attribute objects to their emotions in similar ways to how we attribute objects in the case of humans. In the example of the dog, we say that the dog is happy to see us because its energetic behavior happens after they notice our presence, and because they jump around us. As I argued above, a social model of emotions would be

[60] We can think of the intentionality of emotions by appealing to the distinction drawn by Kenny (1963/2004) between formal and material objects. Formal objects are described abstractly, while material objects are described in concrete terms. For example, when we say that one fears snakes, we must distinguish between the snake described as an instance of danger (formal object), and the snake described as a concrete object (material object). Since different concrete objects can be described under a common abstract description, we can say that two instances of an emotion have the same formal object (danger) but different material objects (snakes or, say, spiders). The upshot of appealing to functional kinds is that formal objects can be described functionally. Hence, we open the door for an account of the intentionality of emotions. An account in this direction was offered previously by Prinz (2004).

incompatible with this description insofar as such this would require sharing a social world with animals, a claim that is at the very least implausible.

Lastly, functional models can make sense of the application of emotion concepts in both social and non-social contexts. This is because functional descriptions can include situational elements which may involve social practices, but do not require them from the outset. Hence, from a functionalist perspective, we can spell out emotions that either require or do not require social practices. This makes functional frameworks more generalizable than social ones.

In sum, functionalism offer the best of the models above while avoiding their most pressing pitfalls. From essentalist and historical models, functionalism allows accounting for embodiment and the presence of neural and physiological mechanisms by taking them to be the realizers of functional profiles. In contrast to these models, however, functionalism does not demand strict mappings that preclude variability. From social models, functionalism can also integrate social aspects of the more complex emotions, while enabling descriptive accounts of the embodiment of emotions, making sense of our attribution to non-human animals, and avoiding confusions that stem from the intentionality of emotions.

In what is left, I will address some of the most salient objections to functionalism coming from both general discussions in philosophy of mind and from local ones regarding emotions proper. After I tackle these objections, I will draw some consequences of this strategy and come back to the main question of this work: how can we offer a general framework for a scientific theory of emotions and how can it deal with empirical demands.

6.3 Objections to functionalism

6.3.1 Functionalism is not falsifiable

One common objection against functionalism is that it cannot be falsified in principle. The main reason for this, skeptics claim, is that we can describe anything functionally. Hence, no amount of empirical evidence can falsify a candidate functional description. This makes functionalist frameworks scientifically unattractive, according to these skeptics.

This objection was framed by Churchland (1981) against functionalism in general. In his presentation, he invites us to consider a functional description of alchemical kinds such as the “spirit mercury." Substances that were “ensouled by mercury" could be characterized functionally in terms of their disposition to reflect light or liquify under heat. If positing functional kinds were a promising scientific strategy, there would be then good reasons not to eliminate these kinds, Churchland argues.

A similar argument can be put forward using the case of phlogiston and other eliminable kinds. In Churchland’s words:

> The alchemical example is a deliberately transparent case of what might well be called “the functionalist strategem," and other cases are easy to

> imagine. A cracking good defense of the phlogiston theory of combustion can also be constructed along these lines. Construe being highly phlogisticated and being dephlogisticated as functional states defined by certain syndromes of causal dispositions; point to the great variety of natural substrates capable of combustion and calxification; claim an irreducible functional integrity for what has proved to lack any natural integrity; and bury the remaining defects under a pledge to contrive improvements. A similar recipe will provide new life for the four humors of medieval medicine, for the vital essence or archeus of pre-modern biology, and so forth. (Churchland, 1981, p. 81)

The functionalist strategem, as Churchland calls it, cannot save these kinds from elimination. In Churchland's case, it cannot save folk-psychology. Applied to our case, it cannot save emotions.[61]

The application of this argument to emotions has also been put forward by Barrett (2016). She claims that under the functionalist framework, definitions are stipulated, not discovered. She argues that an emotion can be associated with a variety of functions. If this is so, we can ask: how can we find the *correct* functional profile that corresponds to an emotion? According to Barrett, this can only be done by stipulation. As a result, claims about emotions corresponding to functional profiles are not empirical claims, but conceptual ones that are immune to falsification.

This line of argument depends on the premise that functional characterizations are always stipulated and cannot be subject to empirical investigation. According to this objection, appealing to a functional profile is an ad hoc move that cannot lead to a progressive scientific research program. Yet, I believe this view conflates two kinds of functionalism.[62] They are what Block (1978) called Functionalism (with a capital F, also called Analytic Functionalism) and Psychofunctionalism:

> One can [...] categorize functionalists in terms of whether they regard functional identities as part of a priori psychology or empirical psychology. The a priori functionalists (e.g., Smart, Armstrong, Lewis, Shoemaker) are the heirs of the logical behaviorists. They tend to regard functional analyses as analyses of the meanings of mental terms, whereas the empirical functionalists (e.g., Fodor, Putnam, Harman) regard functional analyses as substantive scientific hypotheses. (Block, 1978, p. 67)

For Analytic Functionalists, as Block explains, a functional characterization is indeed an a priori truth. It is a matter of logical analysis whether an object satisfies a

[61] There is an important difference between folk-psychological kinds, or at least emotions, and the kinds that Churchland invokes in his examples. Churchland appeals to kinds that are arguably candidates to chemical kinds. Yet, the sort of inferences done in chemistry rely on microstructures, hence on a presumably essentialist model of kindhood. In other words, in chemistry, inferences are projected in virtue of concrete shared structures rather than functional profiles. This is why it is possible to eliminate these kinds from chemistry. As I argued before, I do not think this is the case of folk-psychological kinds, or at least of emotions.

[62] Churchland acknowledges this distinction, but does not consider it in his argument.

functional description. This is the case of some of the functional kinds I have invoked before, such as levers. Describing a lever as an object that allows force transfer in a specific way is a matter of analysis, not of empirical fact. In other words, the functional description specifies what it *means* for something to be a lever, not a fact about these objects.

Yet, not all functional kinds are stipulated. Consider the kind "photosynthetic organism," a central kind in biology. Photosynthetic organisms are quite varied, even carrying out photosynthesis in a myriad of ways. Still, the kind of photosynthetic organisms is a functional kind. It is the kind that consists of those organisms that carry out photosynthesis, however that may be. This kind is not merely stipulative, as it is a kind discovered by biology. Furthermore, membership to the kind does not happen by an act of stipulation, but as a matter of empirical fact. In other words, whether or not an organism is a photosynthetic organism is not part of what defines these organisms. Rather, organisms that count as photosynthetic organisms count as such in virtue of possessing means to carry out photosynthesis, even if these are not realized in the same way in different members of the kind, and whether or not an organisms has such means is a matter of empirical fact. If this is true, it follows that attributing a functional profile can be a scientific hypothesis, not a stipulation.

This kind of process is what Psychofunctionalists have in mind when they claim that mental states can be cashed out functionally, and that I have in mind when I claim that emotions can be constructed scientifically as functional kinds. For Psychofunctionalists, it is up to empirical psychology to find the best functional characterization of the phenomena it is interested in. In my view, this applies to emotions. When we claim that fear is a quick reaction to perceived danger, we can submit that to empirical investigation. On my account, this happens in two ways.

First, recall the process of reconstitution as explication developed in chapter 5. This process called for an analysis of folk concepts which then would be explicated in a target vocabulary, in this case a functional vocabulary. While this may seem as a task of pure conceptual analysis of folk concepts, this analysis can also be empirically informed with regards to how we attribute emotions to others and to ourselves. In the first part of this chapter, I presented a number of experiments in psychology which brought out a number of features that help us pinpoint the phenomenon we intend to explain. All these experiments relied on subjects' pretheoretical concepts to attribute emotions to others and themselves, and showed that emotional attribution depends on bodily and expressive cues, contextual factors, and so on. It is important to notice that these facts can change in different historical moments and across different cultures, and constitute a contingent basis on which we can carry out subsequent exercises of explication and theory construction. Insofar as our folk concepts change through time, it is a matter of ongoing investigation how these changes occur and how they change the ways in which we attribute emotions and hence make reference to the phenomenon to be explained.[63]

[63] It might be possible to open space for experimental philosophy approaches in this regard as well. I will not pursue that at the moment. For the time being, I submit that many of the

Second, whether or not the candidate functional characterization adequately captures the phenomenon identified through the aforementioned diagnostic features is also a matter of empirical investigation. We could in principle find instances of an emotion that do not fit the candidate functional characterization, leading to a refinement of the candidate altogether. For example, under Lazarus's core relational themes presented above, sadness is characterized as the experience of an irrevocable loss. By looking at instances of sadness in depression, for instance, we could find out that this is not an adequate characterization, as there may be instances of sadness that do not arise from a perceived irrevocable loss. If we confirm this, we must look for a different way of cashing out sadness functionally. Again, this can be done through empirical means by investigating how well the candidate characterization fits actual instances of the emotion. As a result, whether or not a functional description captures a given emotion is a hypothesis, not a stipulation.

When we consider that functional descriptions can also be quite complex, the empirical nature of the functionalist project becomes even clearer. As I presented it above, a functional description of emotions in my view must consider three aspects: organism-environment relations (situations), relations to other inner states, and behavioral dispositions. On the account I am offering, the task for a science of emotions is to find how these three parts fit together for each emotion. We must therefore investigate which situations lead to specific emotions, what other inner states are required for that emotion to obtain, and what description fits the organism's reactions and dispositions. From this perspective we can formulate a number of testable hypotheses, and allow for empirical research to offer the best account.

6.3.2 Functionalism is teleological

Another objection put forward by Barrett (2016) is that functionalism appeals to teleology. This objection is directed towards a specific functional model of emotions, namely, the one proposed by Adolphs (2016, see also Adolphs and Andler 2018). Yet, it is an objection worth considering, as it introduces an important caveat for functional models overall.

In Adolphs's view, emotions are biological functional states that regulate complex behavior both in humans and animals. On this account, emotions serve the function of coping with environmental changes in a flexible, predictive, and context-sensitive manner. This appeal to functions as goals or benefits in our evolutionary history is what is known as a teleological account of functions.

Against Adolphs, Barrett claims that teleology is an act of mental inference, not of identification of a phenomenon. When we attribute a teleological function to an emotion, or any other psychological phenomenon for that matter, we are inferring what that emotion or phenomenon is adaptive for, rather than describing the actual phenomenon itself. As I understand Barrett, she argues that a science of emotions must deal with what she calls "action identification" rather than inference. She contrasts

questions that we can approach with experimental philosophical methods can be approached with psychological methods as well.

the claim that "eyes widen in fear to increase vigilance" with "eyes widen to expand peripheral vision." The first claim, a teleological claim, makes an inference about the purpose of the eyes widening, but it does not describe the action that leads to the satisfaction of that purpose. The second claim does describe the action, without any reference to purposes. It is merely the claim that eyes widen and as a result, peripheral vision is enhanced. Since functionalist models such as Adolphs's make claims of the first sort, instead of the second sort, they fail as models that identify what is going on and instead constitute models of a speculative nature.

Barrett's presentation of the argument is quite unclear. First, the assumption that science must only deal with action identification and not with what she calls mental inferences strikes as implausible. This is because "mental inferences" are quite ambiguous here. Barrett relies claims that teleological language is metaphorical, a claim she attributes to Mayr (2004).[64] What she means by "metaphorical" however is not specified. As I read Barrett, what she has in mind is that attributing a teleological function to a phenomenon is speculation. She holds that teleology involves a "metaphorical language that cannot be verified in physical terms" (Barrett, 2016, p. 34), and therefore it involves "mental inferences or attributions [...] of psychological functions."

On a naïve reading, Barrett's objection is weak insofar as all of our scientific hypotheses are inferential in some sense. When we formulate scientific hypotheses, we do not know yet whether the hypotheses are well confirmed. All we know is that given a phenomenon, it is plausible that some other phenomenon is going on. In this sense, criticizing teleology for being speculative in this sense is confusing at best. On a more charitable reading, I believe that what she is objecting is that there are no criteria to determine which teleological claims are true. This is why she appeals to the idea that teleology cannot be verified in physical terms. On this interpretation, all there is is our speculation about what might be the purpose of a given phenomenon. This would mean that teleological vocabulary is unsuited for scientific theorizing and would make her case at least plausible.

Yet, Barrett is confusing two senses of functions that it is vital to keep separate. These are what Godfrey-Smith (1993) calls "Wright functions" and "Cummins functions" (named after Wright (1973) and Cummins (1975)). Wright functions are the sort of functions that appeal to teleology, and that are used in evolutionary biology and other related disciplines. They are initially analyzed in the following terms:

> The function of X is Z *means*
>
> (a) X is there because it does Z,
>
> (b) Z is a consequence (or result) of $X's$ being there. (Wright, 1976, p. 81, cited in Godfrey-Smith 1993, p. 197)

This analysis, as Godfrey-Smith presents it, is meant to account for explanations in evolutionary biology. For example, evolutionary biology ascribes functions in virtue,

[64] As far as I can see in the work Barrett refers to, Mayr does not make any claim about teleology being metaphorical.

not only of their effects, but how those effects explain why the function has been passed on across generations. To say, for instance, that the function of claws is to aid in hunting is useful insofar as the effects of having claws for an organism that hunts explains how that organism has survived and the trait has been passed on. In this sense, the standard interpretation of Wright functions is as answers to 'Why?' questions (Mayr, 1961; Millikan, 1989).

On this reading of functions, functions are indeed teleological as Barrett thinks. In Barrett's example, the claim that the function of the eye widening in fear is to enhance vigilance aspires to be explanatory in that the ascription of said function to the eye widening would presumably explain why organisms have evolved in such a way that their eyes widen in fear, namely, because it has enhanced vigilance in the past and led to better chances of surviving.

Here we can introduce a first counter-argument against Barrett: in spite of Barrett's conviction, it is not clear that this analysis of functions is not scientifically useful. As the example shows, biologists do make use of these patterns of explanation successfully. Moreover, they do not need reduce them to physical vocabulary in order to render them empirically testable. What a biologist would need to show to support their hypothesis that the eye widens in fear to enhance vigilance is to show how increased vigilance aids in the survival and subsequent reproduction of ancestor species. The explanation only requires that we have good evidence that this effect of the widening of the eyes leads to increased survival, explaining why eyes widening has been passed on through generations. This is a matter of empirical fact, not mental inference or speculation, as Barrett would have it.

Yet, let us examine the second sense of functions presented by Godfrey-Smith: Cummins functions. Cummins functions are more akin to what Barrett calls "action identification." They are functions ascribed in terms of how an effect contributes to a more complex capacity or an overall system. For instance, saying that the function of the heart is to pump blood is to ascribe it a role in the overall working of our body as a system. This need not require any story about how that function came to be or aids in the survival of an organism. All that it requires is that we identify how a phenomenon contributes to a system, i.e., identify its actions. According to the standard interpretation, Cummins functions respond to questions of the form 'How?' (Mayr, 1961; Millikan, 1989).

If we can ground functions in this second sense, this is a final blow against Barrett's objection. It is simply not true that functionalism is inherently teleological, since it can appeal to functions in terms of roles and capacities rather than purposes. But even if it were teleological, it is unclear that this would render it unsuitable for science. Many scientific explanations arguably take teleological forms. At best, Barrett would have to offer a stronger argument against teleology in science in general, and offer reasons why functionalism applied to emotions is inherently teleological. These two claims I find highly implausible.

In Barrett's favor though, there is an important caveat that this reply entails. Adolph's particular account, as Scarantino (2018) shows, oscillates between these two

senses of functions. This is a problem since they lead to very different empirical predictions and each account is suited for different types of explanation.[65] As Scarantino puts it:

> [Adolphs and Andler (2018)] cannot have it both ways: either the relevant functional roles emerge from an observation of current capacities and duplicates have our emotions since they share such capacities, or the relevant functional roles emerge exclusively from a past history of selection, but then the current capacities we observe cannot shed light on functional roles. (Scarantino, 2018, p. 203)

What Scarantino correctly points out is that a functional model of emotions must make a decision between teleological and dispositional accounts of functions. As I said before, this is because these accounts lead to different empirical predictions, thus shaping the empirical content of the theory proposed, and because each account of functions has different explanatory demands and offerings.

Making such a decision lies outside the scope of this work, since it requires an in depth discussion of functions. Nevertheless, let me offer arguments in favor of applying an etiological account of functions to emotions rather than a teleological account.

First, in my view, dispositional accounts better capture our folk-psychological concepts than teleological ones. Consider an everyday case of emotional attribution. When we say that someone is afraid, angry, happy or sad, we mean that they have dispositions to entertain certain thoughts and to behave in certain ways. This makes no reference as to the adaptive purpose of the emotion, but to its role in how the subject as a psychological system behaves. This way of talking about emotions functionally appeals then to the emotion's contribution to the system's (i.e. agent's) behavior.

Second, the question about how to individuate emotions scientifically is not the same as the question of why emotions have evolved. Emotions might have evolved for a number of reasons, but, in my view, they are individuated by how they contribute to an agent's overall behavior. In other words, emotions help explain *how* an agent behaves, rather than why it has acquired those forms of behavior. Hence, I submit that Cummins functions are a better account than Wright functions, at least when it comes to cashing out emotions as functional kinds.

Nevertheless, it is worth pointing out that this does not mean that teleology is irrelevant for a science of emotions. As I have claimed before, the question of what adaptive value emotions have is an important question in its own right, and it may call for a teleological answer. To make things clear, all I want to claim is that for the purposes of characterizing emotion kinds, we should apply a causal, Cummins-style account of functions rather than a teleological one. For the purpose of integrating emotions into an evolutionary framework, however, we may go with a teleological account.

[65] See however Neander (2017) for an opposing view to this claim.

6.4 Conclusion

In this chapter, I have applied the methodology sketched in chapter 5 to answer the question set out in chapter 4, namely, what is the best framework of scientific kinds to construct scientific concepts of emotions. I analyzed some of the features figuring in our folk-psychological, pretheoretical account of emotions, and have argued that the best type of model available to construct scientifically meaningful concepts of emotions is a functional framework. This can be understood as an empirically informed type of conceptual analysis.

In the first part of the chapter, I presented arguments based on theoretical remarks and empirical research on emotion attribution suggesting that emotions in our folk-psychology involve features such as arousal, valence, being motivating states, embodiment, phenomenality, intentionality, flexibility, context-sensitivity, and that folk-emotion concepts apply meaningfully to non-human animals. I claimed that some of these features could be cashed out by the four available models of scientific kinds. Specifically, I argued that this is the case for arousal, motivational roles, and valence. I also suggested that phenomenality is not a feature worth considering, given that it is scientifically uninformative. This lead me to restrict the space of features relevant to decide between different models of emotions to embodiment, intentionality, flexibility, context-sensitivity, and the application of emotion concepts to non-human animals.

With these features in mind, I proceeded to analyze the prospects of constructing scientific concepts of emotions that did justice to them. I argued that essentialist models were not suited to capture flexibility and context-sensitivity, as well as having problems with defining essences in general. Against historical models, I suggested that even though they fare better than essentialist models, they are still overly restrictive in the sorts of flexibility and context-sensitivity that they allow from emotion categories. Furthermore, both of these models were not well suited to tackle the social aspects involved in some higher order emotions. This prevents them from being plausible candidates to propose an overarching theory of emotions.

Regarding social models, I discussed three different possibilities following Khalidi's three-fold distinction between different types of social kinds. I argued that only two of these kinds were plausible candidates to cash out emotion concepts, namely, kinds requiring mental attitudes to be in place but not necessarily towards the type, or kinds requiring such attitudes to be in place towards the type but not the tokens. Nonetheless, I discarded these models as prime candidates to construct emotion kinds given the sorts of cognitive sophistication demanded by these models, which prevent a theory of emotions from being integrated with evolutionary biology and psychology. Additionally, I claimed that social models overstate variability, which can be understood in functional terms without the pitfalls of previous social accounts.

This left only functional models on the table, which I suggested could be supported by an account of emotions that integrates situational aspects with behavioral dispositions and relations to other mental states. I then presented Lazarus's and Prinz's individuation of emotions in terms of core-relational themes as a possible model and

example of such an endeavor, and proceeded to present some of the advantages of functional models. Particularly, I stressed the advantage of having a model which allows for multiple realizability, allowing integrating variability, flexibility, and context-sensitivity without sacrificing unity. Additionally, I argued that functional models could capture a range of complexity that would allow proposing an account of emotions that covers basic as well as higher order emotions.

In the last section, I responded to some of the main objections raised against functionalist models in the literature. First, I rejected the claim that functionalism is not falsifiable, recommending a psychofunctionalist approach which takes correspondences between emotion categories and functional descriptions as a matter of empirical hypotheses rather than conceptual truths. This enables functionalist theories to provide scientifically interesting hypotheses which are falsifiable and testable. Second, I discussed whether functionalist accounts must be committed to a teleological account of functions, and argued for an etiological account of function instead. As a result, I concluded that emotions must be understood in terms of the contributions they make to an agent's overall patterns of behavior, being combinations of situational factors, behavioral dispositions, and relations to other mental states.

If these arguments are correct, then we can be optimistic that we can propose a scientifically interesting theory of emotions from a functionalist perspective. In what is left, I will revisit the challenges presented in Part I and explain how my account fits together to respond to these challenges. In closing, I shall also explore some limitations of this approach and avenues for future research.

Chapter 7

Concluding remarks

In this work, I set out to answer the question: how can we construct a scientifically meaningful theory of emotions? In Part I, I offered an analysis of the problem. I divided the problem in two challenges which scientific theories of emotions must face, namely, the Theoretical Challenge and the Empirical Challenge. In chapters 1 and 2, I spelled out these challenges in the following terms:

Theoretical Challenge Provide a systematic theoretical framework that provides empirically testable hypotheses and explains all and only the phenomena under the vernacular term "emotion" under the same explanatory resources and under an overarching generic concept of EMOTION.

Empirical Challenge Provide a scientifically meaningful theoretical framework that establishes correspondences between emotion categories and well-coordinated patterns of neural, physiological, and behavioral responses.

Regarding the Theoretical Challenge, I presented Griffiths's arguments to claim that none of our best theories of emotions could successfully overcome it. I then updated Griffiths's arguments, which relied on an outdated landscape of scientific theories of emotions, and argued that even in this updated version, the challenge had not been overcome. This is because neither Basic Emotion Theories, Appraisal Theories, or Psychological Constructionist theories were apt to provide a sound framework that would capture all and only the phenomena under the term 'emotion' without relying on problematic constructs or rendering emotion taxonomies arbitrary. Hence, a new theory of emotions must be proposed if we are to overcome the Theoretical Challenge.

Concerning the Empirical Challenge, I presented Scarantino's analysis of the Variability Thesis, the claim that is thought to support empirical arguments against the possibility of proposing a scientifically tractable overarching account of emotions. The Variability Thesis, understood as the claim that emotions are naturally heterogeneous phenomena, was divided into two separate theses, namely, the No One-to-One Correspondence Thesis (NOC) and the Low Coordination Thesis (LC). I then analyzed each of these theses and their relations, and argued that previous attempts to draw conclusions from them were flawed in that these theses hide important conceptual ambiguities.

After suggesting that the Variability Thesis is best understood as a disjunction of these two claims, I claimed that even having cleared up some of the logical issues with these theses, an important issue must be resolved before drawing empirical conclusions in support of them, namely, that these theses require an account of what counts as a pattern in the neural, physiological, behavioral, expressive, and phenomenological domains. This led to an examination of each of these domains and their relevance to claims about variability. I argued that only the neural, physiological, and behavioral domains provided relevant sources of evidence for claims about variability. This is because, on one hand, phenomenological and expressive patterns are either scientifically intractable, provide irrelevant evidence to the individuation of emotion kinds (specifically claims about the universality of emotional expressions), or they fall back into behavioral patterns altogether. With this arguments in place, I suggested that the best accounts of patterns for the remaining domains are adopting a pattern assignment strategy in the neural domain, clarifying causal and correlation claims in the case of physiology, and spelling out behavioral patterns in terms of action tendencies or behavioral dispositions.

Before concluding Part I, I integrated the analyses of the Theoretical Challenge and Empirical Challenge offered in the previous chapters in order to evaluate the prospects of overcoming them. I claimed that even though these challenges have not been met yet, we should not opt for an eliminativist strategy. Instead, I suggested that by analyzing the notion of natural kinds present in these arguments and their relation to folk-psychological concepts, we could see why other theoretical resources are still available to propose scientifically meaningful theories of emotions. As a result, I suggested that a revisionist approach could be fruitful in overcoming these challenges, an approach that I sketched out in Part II.

The solution proposed in Part II was divided into three parts. First, I discussed the notion of natural kinds, a central concept in the debates about scientific theories of emotions. I argued that appealing to a unique notion of natural kinds leads to an overly restrictive picture of how science constructs projectible categories. But putting projectibility and entrenchment at the forefront, I argued that we should instead frame questions about kinds in terms of *scientific kinds*, i.e., on which criteria do different scientific disciplines construct projectible category. These criteria can be varied, leading to a pluralistic account of scientific kinds. I identified four types of kinds, namely, essentialist, historical, functional, and social kinds. Having drawn these distinctions, the question concerning emotion kinds then becomes which type of kinds best suits emotions as an explanandum phenomenon.

To tackle this question, I proposed engaging in what has been known in the literature on mechanistic explanations as *reconstituting the phenomena*. Reconstitution occurs when scientists take a step back and recharacterize the phenomenon they intend to explain. I proposed a two-step procedure to reconstitute the phenomena in which scientists first find an ostensive device to make reference to the phenomenon, and then proceed to find the level of analysis that will allow them to construct meaningful projectible kinds.

With this notion of reconstitution at hand, I then argued that in the case of emotions, reconstitution is best thought of in terms of *explicating* folk-emotion concepts. In other words, I claimed that folk-emotion concepts can be used as such ostensive device that allows us to make reference to the phenomenon we wish to recharacterize. This called for a clarification of what the relation between folk concepts and their scientific counterparts should be. Against Strawson's objection against explication, I argued that some degree of mismatch between folk terms and scientific concepts is to be expected, but need not lead to a change of subject. As long as scientific constructs are constrained by their folk counterparts, without them being identical, we can avoid changing the subject while constructing useful scientific concepts. Lastly, I explored Scarantino's claim that folk terms and scientific concepts of emotion should be kept separate, and claimed that in order to avoid changing the subject while producing useful scientific theories, we must be keep maintain a robust degree of similarity between folk emotion concepts and scientific ones while picking scientific concepts by their projectibility, even with some degrees of freedom so as to allow healthy degrees of mismatch.

In the last chapter, I applied these observations about reconstitution and explication to emotion concepts. I presented evidence that folk emotion concepts involve properties that allow us to pick between the different frameworks for scientific kinds presented in chapter 4. Accommodating these features, I submitted, was best done by a functional model in which emotions are understood as relations between inputs, other mental states, and outputs. Specifically, I claimed that emotions should be cashed out in terms of situational aspects, objects and events as inputs, and behavioral dispositions as outputs.

Now, I will consider some consequences of the approach I have been defending in this work. First, I will revisit the challenges presented in Part I and clarify how my account overcomes the Theoretical Challenge while dissolving the Empirical Challenge. This means that it is possible to offer a scientifically interesting theory of emotions, and that we can do that without recourse to specific and consistent correspondences between emotions and neural and physiological patterns. Regarding behavioral patterns, as I will clarify below, correspondences must be taken as empirical hypotheses, so as to prevent emotion taxonomies from falling into conceptual truths. After clarifying the application of these arguments to the challenges at hand, I will then present some final remarks on future avenues of research based on this account.

7.1 Responding to the challenges

7.1.1 Revisiting the Theoretical Challenge

At last, let us go back to the challenges driving our discussion. In Part I, I presented two challenges that the literature on the philosophy, psychology, and neuroscience of emotions have raised regarding the prospects of a scientific theory of emotions. These

are the Theoretical Challenge and the Empirical Challenge. If we adopt a functional model, as I have recommended, how does this help us tackle these challenges?

Let us start with the Theoretical Challenge. As I formulated it, the Theoretical Challenge was to provide a systematic theoretical framework that provides empirically testable hypotheses and explains all phenomena under the vernacular term "emotion" under the same explanatory resources and under an overarching concept of EMOTION. On a functional framework, this challenge can be met.

When presenting and updating the Theoretical Challenge, I claimed that none of the current theories of emotions were able to overcome this challenge. Generally speaking, this was because most contemporary theories of emotions are based on dubious constructs such as "core affect" or "basic emotion," or because they could not explain emotion taxonomies in a non-arbitrary way. Regarding the first problem, proposing a sound theory of emotions will depend on which constructs are invoked, and as I have explained before, I will not attempt to propose a specific theory at the moment. The argument I have been putting forward, however, does say something about the second problem.

In my view, functional characterizations, interpreted as I have suggested, do allow us to cash out emotion taxonomies non-arbitrarily. This is because in the picture I have proposed, what makes emotion categories projectible is their functional unity, which in turn is discovered through empirical means. As a result, emotion taxonomies depend on empirical facts about how we categorize emotions and, most importantly, how these emotions lead to common patterns of behavioral dispositions. Not just any state will count as part of an emotion, and not any functional characterization will be successful. This renders emotion taxonomies scientifically interesting, avoiding the pitfalls that other theories ran into.

Additionally, a functional model of emotions is well-suited to overcome other problems that pertain particular theories of emotions. First, consider the objection against basic emotion theories claiming that these theories are not well-suited to explain higher order emotions such as social or moral emotions. The reason alluded was that basic emotion theories endorsed a very primitive view of emotions which, although well prepared to handle basic cases, could not be projected onto more cognitively sophisticated ones.

On a functional model, however, we need not run into such a problem. Given that functional descriptions allow for complexity, we can offer a description of emotions ranging from simple cases such as fearing snakes to more complex cases such as moral and social emotions. Furthermore, this can be done using the same explanatory resources, since all of the emotions are couched in the same functional terms. In the list of emotions offered above, for example, we can see emotions ranging from fear and anger to pride and shame, emotions with varying degrees of complexity but all individuated in functional terms. Even within an emotion category, we can cash out a range of instances from the simpler to the more complex ones. Consider the char-

acterization of fear as a reaction to facing an immediate and overwhelming danger.[66] Objects satisfying the description of being 'an immediate and overwhelming danger' can be as simple as the presence of a snake, but also as complex as the risk of failing an important job interview. Insofar as these objects can be fit into a common abstract description, we can capture simple and complex instances of the same emotion category in the same functional vocabulary.

Second, concerning appraisal theories, functional accounts also display some advantages. As I explained above, functional accounts do not render emotion taxonomies a matter of conceptual truth, avoiding the issues that were presented against discrete appraisal theories. Additionally, functional descriptions need not lead to the explosion of emotion categories that dimensional appraisal theories led to. On the account I have proposed, emotion categories are limited to those we can construct based on folk-psychological categories without sacrificing projectibility. Given the constraints I have argued for in terms of the possible mismatch between scientific theories and folk-concepts, theories of the sort I envisage do not lead to the counter-intuitive and problematic conclusion that there are infinite types of emotions as dimensional appraisal theorists claim.

Lastly, compared to psychological constructionism, functional theories also fare quite well. Recall that psychological constructionism is based on the idea that emotions are heterogeneous phenomena, as explained in chapters 1 and 2. While psychological constructionists take it as their best suit to accommodate variability from the outset, they do so at the expense of relying on problematic constructs such as "core affect" and at risk of making emotion taxonomies arbitrary. Functional accounts, in contrast, can also accommodate variability without running into these problems.

On one hand, functional descriptions allow for multiple realization, making heterogeneity at the neural and physiological levels unproblematic. On the other hand, however, this need not imply that emotions are acts of categorization dependent on our socially constructed conceptual practices, as Barrett claims. Instead, variability is accommodated without giving up unity, this time spelled out in terms of how an emotion contributes to an agent's overall behavior. Since theories need not give up the conceptual unity of emotion categories in spite of their heterogeneity at the neural and physiological levels, we need not invoke dubious processes of construction to account for emotion taxonomies. What we need is to investigate the commonalities between different instances of emotions in abstract terms and refine our categories accordingly. Hence, we can have the main advantage of constructionist accounts without many of its problems.

In sum, there are good reasons to believe that on a functional framework, we can propose theories that explain the phenomena under the vernacular term 'emotion' using a common set of explanatory resources and under an overarching concept

[66] This characterization differs from the one above in that it does not demand a concrete and physical danger for a fear reaction to obtain. Nevertheless, this is not a problem for the current account, as I have argued that these characterizations should be refined empirically.

of EMOTION. In other words, there are good reasons to be optimistic that we can overcome the Theoretical Challenge by adopting a functionalist approach.

7.1.2 Revisiting the Empirical Challenge

With regards to the Empirical Challenge, there are a number of interesting consequences. The Empirical Challenge was formulated in terms of finding a set of specific and consistent mechanisms that corresponded to each emotion kind. Given the amount of empirical evidence suggesting there were no such mechanisms at the neural and physiological evidence, the prospects of tackling this challenge seemed bleak. As I discussed in chapter 2, meeting this challenge required an analysis of the Variability Thesis, which in turn required an account of how to individuate the candidates for correspondence for each emotion. These are neural and physiological mechanisms, and behavioral patterns.

Adopting a functional framework offers clues as to how to analyze variability. First, with regards to neural and physiological mechanisms, the functional account integrates them into the picture as the realizers of an emotion's functional description. They are part of how we as psychological systems instantiate specific functional patterns. These functional patterns, in turn, are obtained by explicating folk emotion concepts, which make reference to patterns of behavior, into a functional vocabulary.

As a result, we have that we individuate behavioral patterns in terms of a common functional description capturing the commonalities between instances of an emotion in folk-psychological terms. This functional description acts as the explicatum of the folk concept and that is allowed to shift in the presence of empirical demands. Once these functional patterns are individuated (which already enables formulating a number of empirical hypotheses), we can ask how they are instantiated in our brains and bodies. In other words, we can ask how mechanisms at the neural and physiological levels contribute to the overall functional characterization of the emotion.

The upshot of this proposal is that we can now approach variability in clearer terms. On this view, variability is cashed out in terms of multiple realization. To claim that an emotion is a heterogeneous phenomenon at the neural and physiological levels is to say that there are a number of neural and physiological mechanisms realizing a given emotion. Thus, there is no one-to-one correspondence between the functional characterization of an emotion and its realizers, but given that functionalism opens the door for multiple realization, this need not come as a problem for my account. In other words, the fact that we lack correspondences at the neural and physiological level is not evidence against the possibility of proposing scientifically fruitful emotion concepts. As for the Low Coordination thesis, we can spell it out as the claim that these various mechanisms can combine in a number of ways, precluding clear-cut groupings of mechanisms. These claims, of course, remain a matter of empirical investigation.

More interestingly for the overall project, however, is that on the functional framework, the Variability Thesis need not entail elimination. Hence, the Empirical Challenge is dissolved. A theory of emotions on this view need not find specific and consistent mechanisms corresponding to emotion categories. It can integrate various mecha-

nisms as the various realizers of a common functional profile. This not only alleviates worries about the Variability Thesis being true, but makes an interesting and provocative empirical claim. In other words, we can explore the Variability Thesis and the hypotheses that follow from it without giving up our scientific categories of emotions.

Lastly, in my account of the Variability Thesis and the Empirical Challenge, I argued that behavioral patterns are also relevant for claims about correspondence and coordination. I cashed out these behavioral patterns in terms of action tendencies, that is, as behavioral dispositions. This ties in well with the functionalist framework I have defended. On this account, emotions can involve behavioral dispositions as outputs. Furthermore, this yields a number of interesting empirical hypotheses. Given that I have recommended a psychofunctionalist approach, the presumed correspondences between emotions and their functional descriptions (which include these behavioral dispositions) become a matter of empirical fact. This leads to the possibility of offering fruitful and testable theories of emotions.

7.2 Future avenues for a science of emotions

To come to a closing, let us discuss some open questions and avenues for future research. To begin, I have only offered a possible functional characterization of emotions in terms of patterns of dispositional reactions, situations and other mental states. This rough sketch deserves further refinement, which entails both theoretical development and contrasting it with empirical findings. Additionally, other functional accounts deserve attention as rivals to this view. One such model is the aforementioned one proposed by Adolphs (2016).

Besides the two models mentioned, I believe that some traditional theories of emotions can be formulated using functional vocabulary, thus becoming interesting contenders in the arena. Specifically, I believe that basic emotion theories and discrete appraisal theories allow for such transformation. Regarding the former, if we understand basicality in functional terms, we can formulate many of the hypotheses of basic emotion theory using functional kinds. And with regards to the latter, cashing out syndromes in functional terms and neural and physiological components as realizers, we can rework the theory to fit it into the functional framework.

Lastly, the functional account of emotion kinds invites a revision of existing empirical evidence. We already know a great deal about how emotions are related to our brains, our bodies, and their relations to other psychological phenomena. Understanding them as functional kinds thus invites a reinterpretation of these results under the light of functions. For example, we can revisit findings about the different brain mechanisms underlying emotions and redescribe them according to multiple realization. Or we can visit findings about the objects and behavioral outcomes of emotions to construct functional models of each specific emotion. This is to say, a new model need not give up previous findings. On the contrary, it calls for a revision of these findings from a new perspective.

Even though I am optimistic about the prospects of a theory of emotions along the meta-theoretical constraints I have proposed, one issue remains standing, although I will not attempt to solve it here. Given that my account takes folk emotion concepts as ostensive devices that provide the basis for reconstitution and explication, one possible worry is that this will lead to as many theories of emotions as there are different folk psychological frameworks. If different folk psychologies offer different potentially untranslatable folk taxonomies, then the question remains on how to pick the proper taxonomy for a generalizable science of emotions.

With regards to this worry, I will offer two remarks to ameliorate it. First, I believe that we can find a folk taxonomy that is applicable universally if we keep in mind that functional descriptions can be quite abstract. By allowing ourselves these degrees of abstraction, we can approach the most general taxonomy possible, plausibly leading to a generalizable theory.

Nevertheless, we cannot guarantee that this will always be possible. In this case, I believe we must accept certain amount of relativity with regards to our scientific theories of emotions. If we cannot offer a universal taxonomy on which to base the construction of our theories, then we must accept that our theories apply to a specific set of phenomena indicated by the taxonomies we pick. However, I do not believe this to be an argument against a science of emotions. On one hand, even if we could pinpoint a universal folk-psychological framework, we should always expect shifts as science progresses and influences folk concepts. What this means is not that a generalizable theory is impossible, but that our theories will keep changing as science progresses, a consequence that I find desirable, as it least to progressive research programs.

On the other hand even if the chosen folk-psychological framework is relative to a specific group, this need not mean that a theory of emotions along those line is not an interesting theory. We could still find interesting facts about how we, whichever group that makes reference to, live our emotional lives. It is then a matter of investigation to what extent these theories can be generalized. Given the framework I have presented, I believe there are good reasons to be optimistic.

In sum, I have offered a meta-theoretical account of how to construct scientific theories of emotions. This account uses a functional framework to construct projectible scientific kinds based on the explication of folk emotion concepts, and allows us to overcome the challenges that previous theories of emotions have faced. What is left is to propose theories along these lines in order to guide scientific investigation and hopefully find out what emotions are. This is a task that only future research will validate. What I have done is provide a ground on which to stand and start inquiring once again.

References

Adolphs, R. (2016). How should neuroscience study emotions? by distinguishing emotion states, concepts, and experiences. *Social Cognitive and Affective Neuroscience*, *49*(4), 24–31.

Adolphs, R., & Andler, D. (2018). Investigating Emotions as Functional States Distinct From Feelings. *Emotion Review*, *10*(3), 191-201.

Anderson, M. L. (2015). Mining the Brain for a New Taxonomy of the Mind. *Philosophy Compass*, *10*(1), 68-77.

Aristotle. (n.d./2009). *The Nicomachean ethics* (D. Ross, Trans.). Oxford: Oxford University Press.

Aristotle. (n.d./2018). *The art of rhetoric* (R. Waterfield, Trans.). Oxford: Oxford University Press.

Averill, J. R. (1980). A constructivist view of emotion. In R. Plutchik & H. Kellerman (Eds.), *Emotion, theory, research, and experience* (Vol. 1: Theories of emotion, p. 305-339). New York: Academic Press.

Baars, B. J., & Gage, N. M. (2010). *Cognition, brain, and consciousness: Introduction to cognitive neuroscience* (2nd ed ed.). Burlington, MA: Academic Press/Elsevier.

Bach, T. (2012). Gender Is a Natural Kind with a Historical Essence. *Ethics*, *122*(2), 231–272.

Baker, L. R. (1993). Eliminativism and an Argument from Science. *Mind & Language*, *8*(2), 180-188.

Bar-On, D. (2017). Communicative Intentions, Expressive Communication, and Origins of Meaning. In K. Andrews & J. Beck (Eds.), *The Routledge handbook of philosophy of animal minds* (p. 301-312). New York: Routledge.

Barrett, L. F. (2006). Are Emotions Natural Kinds? *Perspectives on psychological science*, *1*(1), 28–58.

Barrett, L. F. (2012). Emotions are real. *Emotion*, *12*(3), 413–429.

Barrett, L. F. (2014). The Conceptual Act Theory: A Precis. *Emotion Review*, *6*(4), 292–297.

Barrett, L. F. (2016). Functionalism cannot save the classical view of emotion. *Social Cognitive and Affective Neuroscience*, *12*(1), 34-36.

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*(1), 1-23.

Barrett, L. F. (2018a). *How Emotions Are Made.* London: Pan Books.

Barrett, L. F. (2018b). *Yes. when I was writing my book for the public ("How Emotions are Made"), one of my editors complained (rightly) that "Conceptual Act" didn't sound like a theory of emotion, so I changed it to be more specific & self-explanatory. [Tweet].* Retrieved from `https://twitter.com/LFeldmanBarrett/status/1047256606617026560`

Barrett, L. F., Lindquist, K. A., Bliss-Moreau, E., Duncan, S., Gendron, M., Mize, J., & Brennan, L. (2007). Of Mice and Men: Natural Kinds of Emotions in the Mammalian Brain? A Response to Panksepp and Izard. *Perspectives on Psychological Science*, *2*(3), 297–312.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 421-441.

Bechtel, W., & Richardson, R. C. (2000/2010). *Discovering complexity: Decomposition and localization as strategies in scientific research.* Cambridge, MA: MIT Press.

Bell, C. (1844). *The anatomy and philosophy of Expression as connected with the fine arts.* London: John Murray.

Block, N. (1978). Troubles with Functionalism. *Minnesota Studies in the Philosophy of Science*, *9*, 261–325.

Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies*, *61*, 127–148.

Boyd, R. (1999a). Homeostasis, Species, and Higher Taxa. In R. A. Wilson (Ed.), *Species: New interdisciplinary essays* (pp. 141–185). Cambridge, MA: MIT Press.

Boyd, R. (1999b). Kind, Complexity and Multiple Realization. *Philosophical Studies*, *95*, 67–98.

Boyd, R. (2010). Realism, Natural Kinds, and Philosophical Methods. In H. Beebee & N. Sabbarton-Leary (Eds.), *The semantics and metaphysics of natural kinds* (pp. 212–234). New York: Routledge.

Brown, A. S., & Marsh, E. J. (2008). Evoking false beliefs about autobiographical experience. *Psychonomic Bulletin & Review*, *15*(1), 186-190.

Carnap, R. (1950/1963). *Logical Foundations of Probability.* Chicago: The University of Chicago Press.

Chakravartty, A. (2007). *A Metaphysics of Scientific Realism.* Cambridge: Cambridge University Press.

Chalmers, D. J. (1997). *The conscious mind: In search of a fundamental theory.* New York: Oxford Univ. Press.

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes. *The Journal of Philosophy*, *78*(2), 67-90.

Churchland, P. M. (1985). Conceptual Progress and Word/World Relations: In Search of the Essence of Natural Kinds. *Canadian Journal of Philosophy*, *15*(1), 1–17.

Clark-Polner, E., Johnson, T. D., & Barrett, L. F. (2017). Multivoxel Pattern Analysis Does Not Provide Evidence to Support the Existence of Basic Emotions. *Cerebral Cortex*, *27*(3), 1944-1948.

Collet, C., Vernet-Maury, E., Delhomme, G., & Dittmar, A. (1997). Autonomic nervous system response patterns specificity to basic emotions. *Journal of the Autonomic Nervous System*, *62*(1), 45-57.

Colombetti, G. (2009). From affect programs to dynamical discrete emotions. *Philosophical Psychology*, *22*(4), 407–425.

Colombetti, G. (2017). *The feeling body: Affective science meets the enactive mind.* Cambridge, MA: MIT Press.

Conway, M. A., & Bekerian, D. A. (1987). Situational knowledge and emotions. *Cognition and Emotion*, *1*(2), 145-191.

Cooper, J., Zanna, M. P., & Taves, P. A. (1978). Arousal as a necessary condition for attitude change following induced compliance. *Journal of Personality and Social Psychology*, *36*(10), 1101-1106.

Cordaro, D. T., Sun, R., Keltner, D., Kamble, S., Huddar, N., & McNeil, G. (2018). Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion*, *18*(1), 75-93.

Cramer, A. O. J., van Borkulo, C. D., Giltay, E. J., van der Maas, H. L. J., Kendler, K. S., Scheffer, M., & Borsboom, D. (2016). Major Depression as a Complex Dynamic System. *PLOS ONE*, *11*(12), e0167490.

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, *153*(3), 355–376.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience.* Oxford: Oxford University Press.

Craver, C. F. (2009). Mechanisms and natural kinds. *Philosophical Psychology*, *22*(5), 575–594.

Craver, C. F., & Tabery, J. (2017). Mechanisms in Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 ed.). Metaphysics Research Lab, Stanford University.

Crivelli, C., & Fridlund, A. J. (2018). Facial Displays Are Tools for Social Influence. *Trends in Cognitive Sciences*, *22*(5), 388-399.

Crozier, W. R. (2014). Differentiating Shame from Embarrassment. *Emotion Review*, *6*(3), 269-276.

Cummins, R. (1975). Functional Analysis. *The Journal of Philosophy*, *72*(20), 741-765.

Darden, L., & Maull, N. (1977). Interfield Theories. *Philosophy of Science*, *44*(1), 43-64.

Darwin, C. (1872/2009). *The expression of the emotions in man and animals* (4th ed., 200th anniversary ed.; P. Ekman, Ed.). Oxford: Oxford University Press.

Davitz, J. R. (1969). *The Language of Emotion.* New York: Academic Press.

DeLancey, C. (2002). *Passionate Engines: What Emotions Reveal About Mind and Artificial Intelligence.* Oxford: Oxford University Press.

Descartes, R. (1649/1985). The Passions of the Soul. In J. Cottingham, R. Stoothoff, & D. Murdoch (Trans.), *The Philosophical Writings of Descartes* (Vol. I, p. 111-151). Cambridge: Cambridge University Press.

de Sousa, R. (1984). The Natural Shiftiness of Natural Kinds. *Canadian Journal of Philosophy*, *14*(4), 561–580.

Dupré, J. (1981). Natural Kinds and Biological Taxa. *The Philosophical Review*, *90*(1), 66.

Dutilh Novaes, C., & Reck, E. (2017). Carnapian explication, formalisms as cognitive tools, and the paradox of adequate formalization. *Synthese*, *194*(1), 195–215.

Edelson, M., Sharot, T., Dolan, R. J., & Dudai, Y. (2011). Following the Crowd: Brain Substrates of Long-Term Memory Conformity. *Science*, *333*(6038), 108-111.

Eickers, G., Loaiza, J. R., & Prinz, J. (2017). Embodiment, Context-Sensitivity, and Discrete Emotions: A Response to Moors. *Psychological Inquiry*, *28*(1), 31-38.

Ekman, P. (1972). Universals and Cultural Differences in Facial Expressions of Emotion. In *Nebraska symposium on motivation* (Vol. 19, p. 207-282). Lincoln: University of Nebraska Press.

Ekman, P. (1980). Biological and cultural contributions to body and facial movement in the expression of emotions. In A. Rorty (Ed.), *Explaining emotions* (p. 73-102). Berkeley: University of California Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3), 169–200.

Ekman, P. (1996/2009). Afterword: Universality of Emotional Expression? A Personal History of the Dispute. In P. Ekman (Ed.), *The expression of the emotions in man and animals* (4th ed., 200th anniversary ed ed., p. 363-393). Oxford: Oxford University Press.

Ekman, P. (2009). Darwin's contributions to our understanding of emotional expressions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3449-3451.

Ekman, P., & Cordaro, D. (2011). What is Meant by Calling Emotions Basic. *Emotion Review*, *3*(4), 364–370.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, *17*(2), 124–129.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., ... Masatoshi, T. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, *53*(4), 712–717.

Ekman, P., Friesen, W. V., & Tomkins, S. S. (1971). Facial Affect Scoring Technique: A First Validity Study. *Semiotica*, *3*(1).

Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science*, *221*(4616), 1208–1210.

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-Cultural Elements in Facial Displays of Emotion. *Science*, *164*(3875), 86-88.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, *128*(2), 203-235.

Ereshefsky, M., & Matthen, M. (2005). Taxonomy, Polymorphism, and History: An Introduction to Population Structure Theory. *Philosophy of Science*, *72*(1), 1–21.

Ereshefsky, M., & Reydon, T. A. C. (2015). Scientific kinds. *Philosophical Studies*, *172*(4), 969–986.

Fehr, B., & Russell, J. A. (1984). Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General*, *113*(3), 464-486.

Fine, A. I. (1984). The Natural Ontological Attitude. In J. Leplin (Ed.), *Scientific Realism* (pp. 261–277). Berkeley: University of California Press.

Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, *28*(2), 97–115.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology.* Cambridge, MA: MIT Press.

Frankel, A. (2016). *"Do you see how much I'm suffering here?": Abuse against transgender women in US immigration detention.* New York, N.Y.: Human Rights Watch.

Frijda, N. H. (1988). The laws of emotion. *American Psychologist*, *43*(5), 349-358.

Frijda, N. H. (2007). *The laws of emotion.* Mahwah, NJ: Erlbaum.

Frijda, N. H., Kuipers, P., & ter Schure, E. (1989). Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology*, *57*(2), 212-228.

Frijda, N. H., & Zeelenberg, M. (2001). Appraisal: What is the dependent? In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (p. 141-155). Oxford: Oxford University Press.

Gazzaniga, M. S., & Mangun, G. R. (Eds.). (2014). *The cognitive neurosciences* (Firth edition ed.). Cambridge, MA: MIT Press.

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014). Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion*, *14*(2), 251-262.

Glennan, S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, *44*(1).

Glennan, S. (2002). Rethinking Mechanistic Explanation. *Philosophy of Science*, *69*(S3), S342-S353.

Godfrey-Smith, P. (1993). Functions: Consensus without unity. *Pacific Philosophical Quarterly*, *74*(3), 196-208.

Godman, M. (2013). Psychiatric Disorders qua Natural Kinds: The Case of the "Apathetic Children". *Biological Theory*, *7*(2), 144–152.

Goodman, N. (1979/1983). *Fact, Fiction, and Forecast* (4th ed.). Cambridge, MA: Harvard University Press.

Griffiths, P. E. (1994). Cladistic Classification and Functional Explanation. *Philosophy of Science*, *61*(2), 206–227.

Griffiths, P. E. (1997). *What Emotions Really Are: The Problem of Psychological Categories*. Chicago: University of Chicago Press.

Griffiths, P. E. (1999). Squaring the Circle: Natural Kinds with Historial Essences. In R. A. Wilson (Ed.), *Species: New interdisciplinary essays* (pp. 209–228). Cambridge, MA: MIT Press.

Griffiths, P. E., & Stotz, K. (2013). *Genetics and philosophy: An introduction*. Cambridge: Cambridge University Press.

Guala, F. (2014). On the Nature of Social Kinds. In M. Gallotti & J. Michaels (Eds.), *Perspectives on Social Ontology and Social Cognition* (pp. 57–68). Dordrecht: Springer.

Hacking, I. (1991). A tradition of natural kinds. *Philosophical Studies*, *61*, 109–126.

Hacking, I. (1994). Entrenchment. In D. F. Stalker (Ed.), *Grue! the new riddle of induction*. Chicago: Open Court.

Hacking, I. (2007). Natural Kinds: Rosy Dawn, Scholastic Twilight. *Royal Institute of Philosophy Supplement*, *61*, 203–239.

Hareli, S., Elkabetz, S., & Hess, U. (2018). Drawing inferences from emotion expressions: The role of situative informativeness and context. *Emotion*.

Hietanen, J. K., Glerean, E., Hari, R., & Nummenmaa, L. (2016). Bodily maps of emotions across child development. *Developmental Science*, *19*(6), 1111-1118.

Hull, D. L. (1978). A Matter of Individuality. *Philosophy of Science*, *45*(3), 335–360.

Hume, D. (1738/2000). *A treatise of human nature* (D. F. Norton & M. J. Norton, Eds.). Oxford: Oxford University Press.

Hutto, D. D. (2012). Truly Enactive Emotion. *Emotion Review*, *4*(2), 176-181.

Izard, C. E. (1971). *The face of emotion*. New York: Appleton-Century-Crofts.

Izard, C. E. (2007). Basic Emotions, Natural Kinds, Emotion Schemas, and a New Paradigm. *Perspectives on Psychological Science*, *2*(3), 260–280.

Izard, C. E. (2009). Emotion Theory and Research: Highlights, Unanswered Questions, and Emerging Issues. *Annual Review of Psychology*, *60*(1), 1-25.

Izard, C. E. (2011). Forms and Functions of Emotions: Matters of Emotion–Cognition Interactions. *Emotion Review*, *3*(4), 371-378.

Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, *109*(19), 7241-7244.

Jack, R. E., Sun, W., Delis, I., Garrod, O. G. B., & Schyns, P. G. (2016). Four not six: Revealing culturally common facial expressions of emotion. *Journal of Experimental Psychology: General*, *145*(6), 708-730.

James, W. (1884). What is an Emotion? *Mind*, *9*(34), 188–205.

Kassam, K. S., Markey, A. R., Cherkassky, V. L., Loewenstein, G., & Just, M. A. (2013). Identifying Emotions on the Basis of Neural Activation. *PLOS ONE*, *8*(6), e66032.

Kenny, A. (1963/2004). *Action, Emotion and Will*. London: Routledge.

Khalidi, M. A. (2013). *Natural Categories and Human Kinds: Classification in the Natural and Social Sciences*. Cambridge: Cambridge University Press.

Kim, J. (1992). Multiple Realization and the Metaphysics of Reduction. *Philosophy and Phenomenological Research*, *52*(1), 1-26.

Klein, C. (2013). Multiple realizability and the semantic view of theories. *Philosophical Studies*, *163*(3), 683–695.

Kragel, P. A., & LaBar, K. S. (2015). Multivariate neural biomarkers of emotional states are categorically distinct. *Social Cognitive and Affective Neuroscience*, *10*(11), 1437-1448.

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, *84*(3), 394-421.

Kripke, S. A. (1972). *Naming and necessity.* Cambridge, MA: Harvard University Press.

Kronfeldner, M. (2015). Reconstituting Phenomena. In U. Mäki, I. Votsis, S. Ruphy, & G. Schurz (Eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki* (Vol. 1, p. 169-181). Cham: Springer International Publishing.

Kuppens, P., Van Mechelen, I., Smits, D. J. M., De Boeck, P., & Ceulemans, E. (2007). Individual differences in patterns of appraisal and anger experience. *Cognition & Emotion*, *21*(4), 689-713.

Lambie, J. A., & Marcel, A. J. (2002). Consciousness and the varieties of emotion experience: A theoretical framework. *Psychological Review*, *109*(2), 219-259.

Langer, S. K. (1967). *Mind: An Essay on Human Feeling.* Baltimore: The John Hopkins Press.

Lazarus, R. S. (1991). *Emotion and adaptation.* New York: Oxford University Press.

LeDoux, J. E. (2000). Emotion Circuits in the Brain. *Annual Review of Neuroscience*, *23*(1), 155–184.

LeDoux, J. E. (2003). The emotional brain, fear, and the amygdala. *Cellular and Molecular Neurobiology*, *23*(4-5), 727–738.

LeDoux, J. E. (2007). The amygdala. *Current Biology*, *17*(20), 868–874.

LeDoux, J. E. (2012). Rethinking the Emotional Brain. *Neuron*, *73*(4), 653–676.

LeDoux, J. E. (2013). The slippery slope of fear. *Trends in Cognitive Sciences*, *17*(4), 155–156.

LeDoux, J. E., & Brown, R. (2017). A higher-order theory of emotional consciousness. *Proceedings of the National Academy of Sciences*, *114*(10), E2016-E2025.

Levenson, R. W., Ekman, P., & Friesen, W. V. (1990). Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology*, *27*(4), 363–384.

Levine, J. (1983). Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly*, *64*(4), 354-361.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences*, *35*(03), 121–143.

Loaiza, J. R. (2020). Emotions and the problem of variability. *Review of Philosophy and Psychology*.

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, *12*(4), 361-366.

Ludwig, D. (2018). Letting Go of "Natural Kind": Toward a Multidimensional Framework of Nonarbitrary Classification. *Philosophy of Science*, *85*(1), 31–52.

Lutz, C. (1988). *Unnatural emotions: Everyday sentiments on a Micronesian atoll & their challenge to Western theory.* Chicago: University of Chicago Press.

Machamer, P. K., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of science*, *67*(1), 1–25.

Magnus, P. D. (2014). NK≠HPC. *The Philosophical Quarterly*, *64*(256), 471–477.

Magnus, P. D. (2015). John Stuart Mill on Taxonomy and Natural Kinds. *HOPOS: The Journal of the International Society for the History of Philosophy of Science*, *5*(2), 269–280.

Marshall, G. D., & Zimbardo, P. G. (1979). Affective consequences of inadequately explained physiological arousal. *Journal of Personality and Social Psychology*, *37*(6), 970-988.

Mayr, E. (1961). Cause and Effect in Biology. *Science*, *134*(3489), 1501-1506.

Mayr, E. (1982). *The growth of biological thought: Diversity, evolution, and inheritance.* Cambridge, MA: Harvard University Press.

Mayr, E. (2004). *What makes biology unique? Considerations on the autonomy of a scientific discipline.* New York: Cambridge University Press.

Mazzoni, G., & Memon, A. (2003). Imagination can create false autobiographical memories. *Psychological Science*, *14*(2), 186-188.

McCafrey, J. (in press). Does the brain respect basic emotion theory?

McEachrane, M. (2009). Emotion, Meaning, and Appraisal Theory. *Theory & Psychology*, *19*(1), 33-53.

Mendel, G. (1865/2009). Experiments in Plant-Hybridisation. In W. Bateson (Trans.), *Mendel's principles of heredity: A defence, with a translation of Mendel's original papers on hybridisation.* Cambridge: Cambridge Univ. Press.

Michaelian, K. (2010). Is memory a natural kind? *Memory Studies*, *4*(2), 170–189.

Mill, J. S. (1843/1974). *A System of Logic, Ratiocinative and Inductive* (J. M. Robson, Ed.). Toronto: University of Toronto Press.

Millikan, R. G. (1989). An ambiguity in the notion "function". *Biology and Philosophy*, *4*(2), 172-176.

Millikan, R. G. (1999). Historical Kinds and the "Special Sciences". *Philosophical Studies*, *95*(1-2), 45–65.

Moors, A. (2014). Flavors of appraisal theories of emotion. *Emotion Review*, *6*(4), 303–307.

Moors, A. (2017). Integration of Two Skeptical Emotion Theories: Dimensional Appraisal Theory and Russell's Psychological Construction Theory. *Psychological Inquiry*, *28*(1), 1-19.

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal Theories of Emotion: State of the Art and Future Development. *Emotion Review*, *5*(2), 119–124.

Moors, A., & Scherer, K. R. (2013). The role of appraisal in emotion. In M. D. Robinson, E. Watkins, & E. Harmon-Jones (Eds.), *Handbook of cognition and emotion* (p. 135-155). New York, NY: Guilford Press.

Morris, P., Doe, C., & Godsell, E. (2008). Secondary emotions in non-primate species? Behavioural reports and subjective claims by animal owners. *Cognition & Emotion*, *22*(1), 3–20.

Morris, P., Lesley, S., & Knight, S. (2012). Belief in Animal Mind: Does Familiarity with Animals Influence Beliefs about Animal Emotions? *Society & Animals*, *20*(3), 211-224.

Murphy, F. C., Nimmo-Smith, I., & Lawrence, A. D. (2003). Functional neuroanatomy of emotions: A meta-analysis. *Cognitive, Affective, & Behavioral Neuroscience*, *3*(3), 207–233.

Neander, K. (2017). Functional analysis and the species design. *Synthese*, *194*(4), 1147-1168.

Nelson, N. L., & Russell, J. A. (2013). Universality Revisited. *Emotion Review*, *5*(1), 8-15.

Ngo, N., & Isaacowitz, D. M. (2015). Use of context in emotion perception: The role of top-down control, cue type, and perceiver's age. *Emotion*, *15*(3), 292-302.

Nussbaum, M. C. (2001). *Upheavals of Thought: The Intelligence of Emotions*. Cambridge: Cambridge University Press.

Ortony, A., Clore, G. L., & Foss, M. A. (1987). The Referential Structure of the Affective Lexicon. *Cognitive Science*, *11*(3), 341-364.

Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review*, *97*(3), 315–331.

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. Oxford: Oxford Univ. Press.

Panksepp, J. (2007). Neurologizing the Psychology of Affects: How Appraisal-Based Constructivism and Basic Emotion Theory Can Coexist. *Perspectives on Psychological Science*, *2*(3), 281-296.

Panksepp, J. (2008). Carving 'natural' emotions: 'Kindly' from bottom-up but not top-down. *Journal of Theoretical and Philosophical Psychology*, *28*(2), 395-422.

Panksepp, J. (2011). The basic emotional circuits of mammalian brains: Do animals have affective lives? *Neuroscience & Biobehavioral Reviews*, *35*(9), 1791-1804.

Pauen, M. (2017). The Functional Mapping Hypothesis. *Topoi*, *36*(1), 107-118.

Pérez, D. (2013). *Sentir, desear, creer: una aproximación filosófica a los conceptos psicológicos*. Buenos Aires: Prometeo Libros.

Pérez, D. I., & Gomila, A. (2014). La atribución mental y la segunda persona. In T. Balmaceda & K. Pedace (Eds.), *Temas de Filosofía de la Mente* (p. 69-98). Buenos Aires: SADAF.

Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002). Functional Neuroanatomy of Emotion: A Meta-Analysis of Emotion Activation Studies in PET and fMRI. *NeuroImage*, *16*(2), 331–348.

Phelps, E. A., & LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron*, *48*(2), 175–187.

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311.

Plato. (n.d.). Phaedrus. In B. Jowett (Trans.), *The Dialogues of Plato in Five Volumes* (3rd ed., Vol. 1). London: Oxford University Press.

Pöyhönen, S. (2016). Memory as a cognitive kind: Brains, remembering dyads, and exograms. In C. Kendig (Ed.), *Natural kinds and classification in scientific practice* (pp. 145–156). Abingdon, Oxon: Routledge.

Price, C. J., & Friston, K. J. (2005). Functional ontologies for cognition: The systematic definition of structure and function. *Cognitive Neuropsychology*, *22*(3-4), 262-275.

Prinz, J. J. (2004). *Gut reactions. A perceptual theory of emotion.* New York: Oxford University Press.

Prinz, J. J. (2007). *The emotional construction of morals.* Oxford; New York: Oxford University Press.

Putnam, H. (1960/1997). Mind and machines. In *Mind, language and reality* (p. 362-385). Cambridge: Cambridge University Press.

Putnam, H. (1970/1977). Is semantics possible? In S. P. Schwartz (Ed.), *Naming, necessity, and natural kinds* (pp. 102–118). London: Cornell University Press.

Putnam, H. (1975). The meaning of 'meaning'. In *Mind, language and reality: Philosophical papers, volume 2* (pp. 215–271). New York: Cambridge University Press.

Quine, W. V. O. (1969/1977). Natural kinds. In S. P. Schwartz (Ed.), *Naming, necessity, and natural kinds* (pp. 155–175). Ithaca, NY: Cornell University Press.

Rainville, P., Bechara, A., Naqvi, N., & Damasio, A. R. (2006). Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International Journal of Psychophysiology*, *61*(1), 5-18.

Reisenzein, R. (1983). The Schachter Theory of Emotion: Two Decades Later. *Psychological Bulletin*, *94*(2), 239–264.

Remington, N. A., Fabrigar, L. R., & Visser, P. S. (2000). Reexamining the circumplex model of affect. *Journal of Personality and Social Psychology*, *79*(2), 286-300.

Reydon, T. A. C. (2009). How to Fix Kind Membership: A Problem for HPC Theory and a Solution. *Philosophy of Science*, *76*(5), 724–736.

Richards, R. J. (2009). Darwin on mind, morals and emotions. In J. Hodge & G. Radick (Eds.), *The Cambridge companion to Darwin* (2nd ed ed., p. 96-119). Cambridge: Cambridge University Press.

Roca-Royes, S. (2011). Essential Properties and Individual Essences. *Philosophy Compass*, *6*(1), 65–77.

Roseman, I. J. (2011). Emotional Behaviors, Emotivational Goals, Emotion Strategies: Multiple Levels of Organization Integrate Variable and Consistent Responses. *Emotion Review*, *3*(4), 434-443.

Roseman, I. J. (2013). Appraisal in the Emotion System: Coherence in Strategies for Coping. *Emotion Review*, *5*(2), 141-149.

Roseman, I. J., Wiest, C., & Swartz, T. S. (1994). Phenomenology, behaviors, and goals differentiate discrete emotions. *Journal of Personality and Social Psychology*, *67*(2), 206-221.

Roth, M., & Cummins, R. (2017). Neuroscience, Psychology, Reduction, and Functional Analysis. In D. M. Kaplan (Ed.), *Explanation and integration in mind and brain science.* New York, NY: Oxford University Press.

Rottenberg, J., Gross, J. J., & Gotlib, I. H. (2005). Emotion Context Insensitivity in Major Depressive Disorder. *Journal of Abnormal Psychology*, *114*(4), 627-639.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161-1178.

Russell, J. A. (1991). In defense of a prototype approach to emotion concepts. *Journal of Personality and Social Psychology*, *60*(1), 37-47.

Russell, J. A. (1994). Is There Universal Recognition of Emotion From Facial Expression? A Review of the Cross-Cultural Studies. *Psychological Bulletin*, *115*(1), 102-141.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, *110*(1), 145–72.

Russell, J. A. (2009). Emotion, core affect, and psychological construction. *Cognition & Emotion*, *23*(7), 1259–1283.

Russell, J. A. (2015). My Psychological constructionist Perspective, with a focus on conscious affective experience. In L. F. Barrett & J. A. Russell (Eds.), *The Psychological Construction of Emotion* (p. 183-208). New York: The Guilford Press.

Russell, J. A., Lewicka, M., & Niit, T. (1989). A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology*, *57*(5), 848-856.

Saarimäki, H., Ejtehadian, L. F., Glerean, E., Jääskeläinen, I. P., Vuilleumier, P., Sams, M., & Nummenmaa, L. (2018). Distributed affective space represents multiple emotion categories across the human brain. *Social Cognitive and Affective Neuroscience*, *13*(5), 471-482.

Sabini, J., Garvey, B., & Hall, A. L. (2001). Shame and embarrassment revisited. *Personality and Social Psychology Bulletin*, *27*(1), 104–117.

Sabini, J., & Silver, M. (2005). Why Emotion Names and Experiences Don't Neatly Pair. *Psychological Inquiry*, *16*(1), 1-10.

Salmon, N. U. (1981). *Reference and Essence.* Princeton, NJ: Princeton University Press.

Scarantino, A. (2012). How to Define Emotions Scientifically. *Emotion Review*, *4*(4), 358–368.

Scarantino, A. (2015). Basic Emotions, Psychological Construction, and the Problem of Variability. In L. F. Barrett & J. A. Russell (Eds.), *The Psychological Construction of Emotion* (pp. 334–376). New York: The Guilford Press.

Scarantino, A. (2017). Do Emotions Cause Actions, and If So How? *Emotion Review*, *9*(4), 326-334.

Scarantino, A. (2018). Comment: Two Challenges for Adolphs and Andler's Functionalist Theory of Emotions. *Emotion Review*, *10*(3), 202-203.

Scarantino, A., & Griffiths, P. E. (2011). Don't Give Up on Basic Emotions. *Emotion Review*, *3*(4), 444–454.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379–399.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695–729.

Scherer, K. R. (2009a). The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion*, *23*(7), 1307-1351.

Scherer, K. R. (2009b). Emotions are emergent processes: They require a dynamic computational architecture. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1535), 3459-3474.

Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, *66*(2), 310-328.

Schlingloff, L., & Moore, R. (2017). Do Chimpanzees Conform to Social Norms? In K. Andrews & J. Beck (Eds.), *The Routledge handbook of philosophy of animal minds* (p. 381-389). New York: Routledge.

Searle, J. R. (1995). *The construction of social reality.* New York: Free Press.

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, *52*(6), 1061-1086.

Siegel, E. H., Sands, M. K., Van den Noortgate, W., Condon, P., Chang, Y., Dy, J., ... Barrett, L. F. (2018). Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. *Psychological Bulletin*, *144*(4), 343-393.

Slater, M. H. (2015). Natural Kindness. *The British Journal for the Philosophy of Science*, *66*(2), 375–411.

Smedslund, J. (1992). Are Frijda's "Laws of Emotion" Empirical? *Cognition & Emotion*, *6*(6), 435-456.

Solomon, R. C. (2003). *Not Passion's Slave: Emotions and Choice.* Oxford: Oxford University Press.

Spinoza, B. d. (1677/2018). *Ethics: Proved in geometrical order* (M. J. Kisner, Ed. & M. Silverthorne & M. J. Kisner, Trans.). Cambridge: Cambridge University Press.

Spunt, R. P., Ellsworth, E., & Adolphs, R. (2017). The neural basis of understanding the expression of the emotions in man and animals. *Social Cognitive and Affective Neuroscience*, *12*(1), 95-105.

Stieg, C. (2007). Bird Brains and Aggro Apes: Questioning the Use of Animals in the Affect Program Theory of Emotion. *Philosophy of Science*, *74*, 895–905.

Strawson, P. F. (1950). On Referring. *Mind*, *59*(235), 320-344.

Strawson, P. F. (1963). Carnap's Views on Constructed Systems Versus Natural Languages in Analytic Philosophy. In P. A. Schilpp (Ed.), *The Philosophy of Rudolf Carnap* (pp. 503–518). La Salle, IL: Open Court.

Strawson, P. F. (2011). *Introduction to logical theory.* New York, NY: Routledge.

Sullivan, J. (2016). Neuroscientific kinds through the lens of scientific practice. In C. Kendig (Ed.), *Natural kinds and classification in scientific practice* (pp. 47–56). Abingdon, Oxon: Routledge.

Thomasson, A. (2003). Foundation for a Social Ontology. *ProtoSociology*, *18*, 269–290.

Tomkins, S. S. (1962/2008). *Affect, imagery, consciousness* (Vol. I: The Possitive Affects). New York: Springer Pub. Co.

Tomkins, S. S. (1981/1995). The quest for primary motives: Biography and autobiography of an idea. In E. V. Demos (Ed.), *Exploring affect: The selected writings of Silvan S. Tomkins* (p. 27-63). Cambridge: Cambridge University Press & Editions de la Maison des sciences de l'homme.

Touroutoglou, A., Lindquist, K. A., Dickerson, B. C., & Barrett, L. F. (2014). Intrinsic connectivity in the human brain does not reveal networks for 'basic' emotions. *Social Cognitive and Affective Neuroscience*, *10*(9), 1257–1265.

van Fraassen, B. C. (1968). Presupposition, Implication, and Self-Reference:. *Journal of Philosophy*, *65*(5), 136-152.

Vytal, K., & Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: A voxel-based meta-analysis. *Journal of cognitive neuroscience*, *22*(12), 2864–2885.

Weiskopf, D. A. (2011). The Functional Unity of Special Science Kinds. *The British Journal for the Philosophy of Science*, *62*(2), 233–258.

Wilkins, A. M., McCrae, L. S., & McBride, E. A. (2015). Factors affecting the Human Attribution of Emotions toward Animals. *Anthrozoös*, *28*(3), 357–369.

Winkelmayer, R., Exline, R. V., Gottheil, E., & Paredes, A. (1978). The relative accuracy of U.S. British, and Mexican raters in judging the emotional displays of schizophrenic and normal U.S. women. *Journal of Clinical Psychology*, *34*(3), 600-608.

Wittgenstein, L. (1953/2009). *Philosophical investigations* (Rev. 4th ed ed.; P. M. S. Hacker & J. Schulte, Eds. & G. E. M. Anscombe, P. M. S. Hacker, & J. Schulte, Trans.). Malden, MA: Wiley-Blackwell.

Wright, L. (1973). Functions. *The Philosophical Review*, *82*(2), 139-168.

Wright, L. (1976). *Teleological explanations* (First Edition ed.). Berkeley and Los Angeles: University of California Press.

Wundt, W. (1897). *Outlines of Psychology* (C. Hubbard Judd, Trans.). Leipzig: Wilhelm Engelmann.

Yik, M., Russell, J. A., & Steiger, J. H. (2011). A 12-point circumplex structure of core affect. *Emotion*, *11*(4), 705-731.

Selbstständigkeitserklärung
Statement of authorship

Name *Surname*	Loaiza Arias
Vorname *First name*	Juan Raúl
Datum *Date*	24.09.2019
Matrikelnummer *Student ID No.*	

Ich erkläre ausdrücklich, dass es sich bei der von mir eingereichten Arbeit um eine von mir selbstständig und ohne fremde Hilfe verfasste Arbeit handelt.
I expressly declare that the work I have submitted was written independently and without external help.

Ich erkläre ausdrücklich, dass ich sämtliche in der oben genannten Arbeit verwendeten fremden Quellen, auch aus dem Internet (einschließlich Tabellen, Grafiken u.Ä.) als solche kenntlich gemacht habe. Insbesondere bestätige ich, dass ich ausnahmslos sowohl bei wörtlich übernommenen Aussagen bzw. unverändert übernommenen Tabellen, Grafiken o.Ä. (Zitaten) als auch bei in eigenen Worten wiedergegebenen Aussagen bzw. von mir abgewandelten Tabellen, Grafiken o.Ä. anderer Autorinnen und Autoren die Quelle angegeben habe.
I expressly declare that all sources used in the abovementioned work – including those from the Internet (including tables, graphics and suchlike) – have been marked as such. In particular, I declare that I have, without exception, stated the source for any statements quoted verbatim and/or unmodified tables, graphics etc. (i.e. quotations) of other authors.

Mir ist bewusst, dass Verstöße gegen die Grundsätze der Selbstständigkeit als Täuschung betrachtet und entsprechend geahndet werden.
I am aware that violations against the principles of academic independence are considered deception and are punished accordingly.

.....
Unterschrift/*Signature*