# Table of Contents

Abstract .......................................................................................................................... 6
Declaration ......................................................................................................................... 7
Copyright statement ........................................................................................................ 7
Acknowledgements .......................................................................................................... 8
Dedication .......................................................................................................................... 8
Table of figures .................................................................................................................. 9

1 Natural kinds in affective science ............................................................................... 10
   1.1 Overview of main theories ..................................................................................... 10
   1.2 View A: Some categories of individual emotions are natural kinds: Basic Emotion Theory (BET) ........................................................................................................ 13
      1.2.1 Revised versions of BET .................................................................................. 18
   1.3 View B: The category of emotion is a natural kind .................................................. 21
   1.4 View C: The category of emotion is not a natural kind and individual emotions are not natural kinds ......................................................................................... 24
   1.5 Assumptions shared by main views ......................................................................... 25
   1.6 Natural kind status of emotion and empirical science ............................................ 27
      1.6.1 Interpreting experimental data and designing experiments on emotion .............. 29
   1.7 Summing up ............................................................................................................. 35

Part A – Natural kinds and social construction

2 Towards a plausible account of natural kinds ............................................................... 37
   2.1 Metaphysical view vs. epistemological view ........................................................... 37
   2.2 Natural kinds, induction and explanation ............................................................... 38
   2.3 Metaphysical view: Essentialism and natural kinds ............................................... 43
      2.3.1 Discovering essences exclusively a posteriori .................................................. 45
      2.3.2 Excluding kinds without intrinsic essences ....................................................... 47
   2.4 Epistemological view: Property clusters ............................................................... 50
   2.5 Causal Property Clusters (CPCs) and Homeostatic Property Clusters (HPCs) ......... 51
      2.5.1 Property clusters and mechanisms .................................................................... 58
      2.5.2 Property clusters and causal relations ............................................................. 62
   2.6 Stable Property Clusters (SPCs); a more plausible count ...................................... 68
      2.6.1 SPCs and pluralism ......................................................................................... 70
      2.6.2 SPCs, theory dependence and conventionalism ............................................... 72
   2.7 SPCs as a plausible account of natural kinds ........................................................... 76
   2.8 Summing up ............................................................................................................. 77
3 Socially constructed natural kinds

3.1 Social kinds vs. natural kinds

3.2 Objections to the claim that social kinds can be considered natural

3.2.1 Heterogeneity

3.2.2 Interactivity

3.2.3 Normativity

3.3 Replies to objections

3.3.1 On the heterogeneity of social kinds

3.3.2 On the interactivity of social kinds

3.3.3 On the normativity of social kinds

3.4 Social kinds as natural kinds

3.5 Summing up

Part B – The natural kind status of emotion

4 The natural kind status of emotion

4.1 Emotion as a single state vs. emotion as a complex state/process

4.2 BET and the debate on the natural kind status of emotion

4.2.1 BET and defences of the unity of the category of emotion

4.2.2 The universality claim and distinct biological markers

4.2.3 Challenges to the universality claim

4.2.4 Is facial expression necessarily a part of emotion?

4.2.5 Basic emotions and dedicated neural circuits

4.2.6 Challenges to the claim for one-to-one correspondence between basic emotions and dedicated neural circuits

4.3 What is an emotion?

4.3.1 Emotion should not be identified with a neurobiological state lacking a phenomenology

4.3.2 Emotion as a complex process including phenomenological or behavioural features

4.4 What emotion is not

5 Emotion as a social natural kind

5.1 Emotion, induction and explanation

5.2 Emotion as a social kind

5.2.1 Emotion generation and the environment

5.2.2 Emotion generation and passivity

5.3 Emotion as a natural kind

5.3.1 Aptness for induction

5.3.2 Aptness for explanation

5.4 Emotion as a social natural kind

5.5 Implications for the study of affective phenomena
Abstract

In this thesis I argue that the category of emotion is a social natural kind and that this claim has implications for the scientific study of emotion. It is a social kind because it includes social features that are necessary for the explanation of emotional phenomena. It is a natural kind because it is a stable property cluster which shows an aptness for successful induction and explanation.

In chapter 1 I introduce the main views on natural kinds in affective science. I present three different views to claim that all three rely heavily on an influential psychological theory, namely Basic Emotion Theory. I claim that the natural kind status of several affective phenomena is not exclusively a metaphysical or philosophical issue and show its implications for interpreting experimental data and designing experiments.

This thesis is divided in two parts. Part A (chapters 2 and 3) deals with the metaphysics of natural kinds and social kinds. In chapter 2 I outline and reject the traditional view of natural kinds (essentialism) in favour of an epistemological view where natural kinds are property clusters showing an aptness for successful induction and explanation of the phenomena included in them. The epistemological view I endorse has similarities with accounts proposed by Boyd (1991) or Khalidi (2013), however, is primarily based on Slater’s (2015) Stable Property Cluster account. I argue that Slater’s account is a plausible account of natural kinds and has advantages over other available accounts in the literature.

In chapter 3 I argue that we should reject the claim that social kinds cannot be considered natural kinds (or useful scientific categories) because they can be subject to heterogeneous construction, and because they are associated with interactivity and normativity features. I argue that in some cases, heterogeneous construction, interactivity between the clustered properties and the environment or normativity factors, do not necessarily entail that the social kind will not show an aptness for successful induction and explanation. I use the example of gender categories to argue that such social kinds can be stable property clusters and show an aptness for successful induction and explanation.

Part B (chapters 4 and 5) applies the account defended in the first part to affective phenomena and more specifically to the category of emotion. In chapter 4 I claim that emotion is not a neurobiological kind but a social kind; its properties cannot be reduced to neurobiological properties or neurobiological states. I argue that most of the main views presented in chapter 1 rely on BET according to which emotion is a neurobiological kind. I present evidence against this theory and argue that it should not be relied upon to determine the natural kind status of the category of emotion because it oversimplifies, and does not provide a good explanation of, emotional phenomena.

In chapter 5 I argue that emotion is a social kind because it relates to phenomena which are best investigated by looking at several domains in the social sphere such as social psychology, sociology and social ecology. To support my claim I first provide an account of emotion as a stable property cluster which performs a specific function e.g. to enable us to successfully interact with the environment; I then give examples of successful induction and explanation on the basis of which emotion can be considered a natural kind. I conclude that emotion is a social natural kind.


The University of Manchester
Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given the University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licencing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information of the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University’s policy on Presentation of Theses.
Acknowledgements

A thousand thanks are due to my three supervisors: Dr Joel Smith, Professor Helen Beebee and Dr Deborah Talmi. Joel provided invaluable direction, critical insight, challenging comments, guidance, encouragement and support throughout this project. His contribution is evident throughout this thesis (only the good parts). This intellectual labour would not have been delivered without his help and infinite knowledge. My project is a hybrid between two demanding areas of philosophy, namely metaphysics and mind. Hindsight is a wonderful thing (and as Joel of course advised at the beginning) perhaps combining them in a PhD project was a bit extreme. However, I was extremely lucky to have Helen as a second supervisor. She helped me remain sane when things got metaphysically heavy. It can be scary to be supervised by someone who is as breathtakingly brilliant as Helen is. I am grateful to her for being kind, patient and always providing insightful and enabling comments. Deborah provided invaluable knowledge, thought provoking comments and general guidance and support throughout the past three years. She is an inspiration as a scientist and a supervisor and she provided a safe empirical space for me to try out my controversial theoretical ideas.

I am grateful to Pilar Lopez-Cantero for her perpetual support, encouragement and inspiring ideas and for being all around the perfect person to do a PhD with. She provided comments for several parts of this thesis and kept reminding me that life can be fun and rewarding even whilst one is undertaking a PhD. Thanks are also due to Sylvia Barnett for providing support, encouragement and for being there in general.

The wider Manchester Philosophy PhD group 2015-2018 for encouragement, insightful comments and fun times, particularly Abigail Connor, Jack Casey and Aaron Wilson.

Special thanks also are due to Dr Graham Stevens; I benefited from numerous discussions and constructive disagreement on many parts of this thesis.

The further Manchester Philosophy staff/fellows who supported/encouraged my work and offered suggestions at various points: Catharine Abell, Thomas Uebel, Julian Dodd, Chris Daly, David Liggins, Cynthia Macdonald, Graham Macdonald, Jonathan Farrell, Sean Crawford.

My parents Barbara and Nikolaos for being interestingly odd, loving and remarkable individuals and for providing support and encouragement throughout my perpetual metamorphoses. My siblings and their families for all the fun and free therapy they provided over the last three years.

Dedication

This thesis is dedicated to Mema for her patience, love, encouragement and support throughout this project and for keeping the Laestrygonians and the Cyclops at bay.
Table of Figures

Figure 1: The periodic table ................................................................. 39
Figure 2: Types of causal relations ..................................................... 65
Chapter 1- Natural kinds in affective science

In this chapter I give an overview of three main theories on the natural kind status of the category of emotion and of individual emotions. I claim that the three main theories share important assumptions on the metaphysics and nature of emotion, namely that emotion is a neurobiological kind (or that emotions are neurobiological kinds) and that emotion is a natural kind if it is a neurobiological kind. I argue that the debate on the natural kind status of emotion has important practical implications for empirical science by showing that implicit or explicit beliefs held on the nature and natural kind status of emotion influence the interpretation of experimental data and the design of experiments in affective science.

1.1 Overview of main theories

There is an ongoing debate in affective science on the natural kind status of emotion as a category, as well as the natural kind status of categories of individual emotions such as fear or sadness. Some philosophers argue that emotion as a category is a natural kind and that individual categories of emotions such as fear or anger can also be considered natural kinds (Prinz 2004, 102; Charland 2002, 513). Other philosophers such as Griffiths and Scarantino argue that only specific categories of emotions- namely the ones identified as ‘basic emotions’ or the category that includes only ‘basic emotions’- can be considered natural kinds; however, the category of emotion is not a natural kind because it includes heterogeneous phenomena (Griffiths 1997, 16; Scarantino and Griffiths 2011, 452; Scarantino 2012, 392). This debate is reflected in differing views in empirical science; for example Ekman claims that the category of emotion should exclusively include basic emotions (Ekman and Cordaro 2011, 365); his theory allows that if the category includes basic emotions exclusively, it could be considered a natural kind. Barrett and Russell on the other hand, argue that the category of emotion is not a natural kind and that the so called basic emotions should not be considered natural kinds either (Barrett 2006b, 49; Russell 2015, 205).

Broadly speaking the differing views on which categories of emotions are natural kinds reflect a difference in the theories held both on the nature of emotion and also on the

---

1 I use italics to refer to the category of fear or anger rather than to an instance of fear or anger.
metaphysical issue of the existence and characteristics of natural kinds. In the first part of this thesis I deal with the metaphysical issue by suggesting an account of natural kinds building on accounts offered by Boyd (1991), Craver (2009) and Slater (2015). In chapter 2, before discussing the account I suggest we should be using, I briefly outline the traditional view (essentialism) and other popular views, such as pluralism and property cluster views, highlighting the biggest difficulties associated with those accounts. I show that the account I endorse can include a multitude of kinds that are intuitively considered to be natural kinds and are treated as such by scientists. In this way, in chapter 2 I argue for a pluralist understanding of the concept of natural kinds, which can apply to chemical, biological and social kinds. According to the suggested account, natural kinds are categories that permit successful induction and explanation of the phenomena included in them.

The claim that biological kinds can be considered natural kinds has been discussed at length in the last 40 years (Boyd 1991; Griffiths 1997; Hull 1978; Khalidi 2013; Mallon 2016). The claim that social kinds can be considered natural kinds is currently a matter of considerable debate (Haslanger 2005; Khalidi 2013; Mallon 2016). In chapter 3 I investigate whether social kinds in particular can be considered natural kinds by using the example of gender categories, which according to some constructionist views could be considered natural kinds (Haslanger 2005, 11; Mallon 2016, 158). I argue that we should reject views which claim on a priori grounds that social kinds cannot be considered natural kinds, along with associated views according to which social kinds are not useful scientific categories. My claim is not that all social kinds are natural kinds but that some can be, if they meet the conditions set in chapter 2. As suggested by the accounts of Haslanger (2000) and Mallon (2016), gender categories can be considered natural kinds in social ontology, on the grounds that we can create successful generalisations and predictions on the basis of shared external relations between individuals and the environment.

---

2 See Lindquist et al. 2013 for a recent summary of the debate.
3 As I discuss in detail in chapter 3, I take social kinds to be kinds whose objects of investigation fall within the social sphere and whose clusters include several properties which cannot be explained without analysing the relation between the individual and the society.
4 There is a debate on whether gender categories are social categories or whether they are biological categories based on biological sex. In this thesis I accept, but not argue for, the view that these are social categories. See De Beauvoir 1949, Butler 1990, Haslanger 2000 and Mallon 2016 for more detail on this claim. These philosophers provide different accounts on the natural kind status of gender categories but they all agree that these are social and not biological categories.
In the second part of the thesis I focus on the natural kind status of the category of emotion and argue that emotion is a natural kind.\(^5\)\(^6\) In chapter 4, I first discuss and reject the views that emotion is a neurobiological natural kind and that the only emotions that are natural kinds are basic emotions. In chapter 5 I suggest instead that emotion is a social natural kind on the basis of its members sharing the same social functional role. I provide examples of induction and explanation generated by emotion considered as a social kind. According to my view, psychological constructionists such as Barrett (2012) and Russell (2015) are correct in claiming that emotion is not a neurobiological natural kind; instead it is a social kind, it relates to phenomena which are best investigated by looking at several domains in the social sphere such as social psychology, sociobiology and social ecology. However, my view differs from their conclusion that emotion qua social kind cannot be considered a natural kind. Instead, I suggest that emotion as a category is a natural kind because all members of this category perform the same function of enabling a successful social life. My view borrows from Millikan’s (2002) account of biological functions and Greenwood’s (2005) theory of the nature of emotion, which takes emotionality to include continuously changing neural, corporeal and sociocultural resources. In this way my account differs from theories reducing emotion to strictly automatic set responses (e.g. Ekman and Cordaro 2011, 365) or strictly cognitive evaluations (e.g. Solomon 2002, 136). My account also rejects strong constructionist views on emotion, which leave no room for biological processes associated with emotion (e.g. Harré 1986, 4). I conclude that emotion as a social kind can be considered a natural kind and that this claim impacts how we should approach the scientific study of emotion.

The claim that emotion is a social natural kind forms the conclusion of an argument which includes the claims established in chapters 2, 3, 4 and 5 as premises. The argument raised in this thesis has the following form:

P1- Natural kinds are property clusters that show an aptness for successful induction and explanation (epistemological view – chapter 2) regardless of the possession of similar causal mechanisms or the inclusion of interactivity or normativity properties (chapter 3).

P2- Some social kinds can be considered natural kinds according to the epistemological view because they show an aptness for successful induction and explanation (chapter 3).

\(^5\) According to the natural kind account I endorse, specific categories of emotions could also be considered natural kinds however this is a separate claim which I do not argue for in this thesis.

\(^6\) I use italics to refer to the category of emotion rather than to an instance of emotion.
P3- *Emotion* is a social kind because the emotion property cluster includes social properties (chapter 5) that are necessary to explain emotional phenomena. These properties cannot be reduced to physical or neurobiological properties (chapter 4).

P4- *Emotion* is a social kind which shows an aptness for successful generalisation and prediction (chapter 5).

C- *Emotion* is a social natural kind.

In the first part of this chapter I briefly present some of the most dominant views in this debate. In the second part I discuss how this issue influences the science of emotion by affecting the design of experiments and the interpretation of accumulated empirical data. I argue that theories held by researchers on the natural kind status of *emotion* heavily influence the experimental process, even if this issue is not explicitly addressed in the design or interpretation stages. As such, it is very important that researchers on emotion keep up to date with the most plausible theory on the natural kind status of *emotion* so that:

i) They design experiments by explicitly addressing top-down questions about the nature and natural kind status of *emotion*.

ii) They are best able to explain and accommodate accumulated empirical data by integrating, where possible, intradisciplinary and interdisciplinary approaches to the science of emotional phenomena.

1.2 View A: Some categories of individual emotions are natural kinds – Basic Emotion Theory (BET)

The debate on the natural kind status of *emotion* has been greatly influenced by a specific psychological theory on the nature of emotion, namely Basic Emotion Theory (BET). According to this theory certain emotions are sets of automatic responses to specific environmental stimuli; these sets of responses are deemed to be universal, existing in all humans and in some other mammals. In this section I give a brief outline of this theory and then proceed to discuss how it has been integrated within the dominant views on the natural kind status of emotion.
According to BET, specific types of events cause the activation of specific basic emotions. These basic emotions consist of characteristic changes in physiology and behaviour (Ekman 1999, 46-48). To give an example, whenever we are faced with imminent danger (e.g. we see a bear in the wild), our body goes through automatic changes such as changes in blood flow, action preparedness and skin conductivity (changes in physiology). We are disposed to start looking for a place to hide or a direction to run so as not to be noticed by the bear (behavioural changes). According to this theory, changes in physiology are also associated with the expression of emotion. We use our face and body to express specific emotions in specific ways. For example, the emotion of *fear* has been associated with facial expressions such as widening the eyes so as to look for places to hide.

Psychologists who argue for this theory, such as Ekman (1972a) and Izard (2007), focus on certain emotions that are considered to be largely universal such as anger, fear, sadness, happiness, disgust and surprise. This universality is supposed to provide evidence for the claim that a list of specific emotions are pre-wired in humans, that is we are born with them already formed in our brains. For example, ‘the hypothesis is that each basic emotion has a unique neural circuit (or other neural signature)’ (Moors 2009, 645). These circuits are said to be installed by evolution to serve specific adaptational functions. For example, ‘the neural circuit of fear serves survival whereas the neural circuit of anger serves territorial concerns’ (Moors 2009, 645). These neural circuits are neurobiological categories with distinct neurobiological properties, which can be observed by scientific research. According to this theory, the terms ‘hardwired’, ‘basic’ and ‘innate’ are used interchangeably and point to neurobiological markers. These hardwired neurobiological markers (usually viewed as neural circuits) are emotion-specific in that each basic emotion is associated with a distinct neurobiological marker (Scarantino 2015, 363). On what exactly is meant by ‘hardwired’, Scarantino claims that:

---

7 As suggested by Ortony and Turner (1990, 317), the term ‘basic’ has at least three different senses: a conceptual one, a psychological one and a biological one. According to the conceptual sense, basic emotions are theoretically more important than the non-basic emotions. The psychological sense views basic emotions as the building blocks of all other emotional phenomena. This thesis, following Scarantino and Griffiths (2011), focuses on the biological sense, according to which a basic emotion is an emotion with evolutionary origin and distinct biological markers such as a dedicated neural circuit, distinct patterns of Autonomic Nervous System (ANS) activity, distinctive facial signals, etc.
A hardwired circuit is one that is built in the nervous system, inherited, present at birth, and homologous to the brain circuits present in evolutionary related species. (Scarantino 2015, 341)

According to Ekman and Cordaro (2011, 365) the main characteristics found in most (nearly all) basic emotions are the following:

1. Distinctive universal signals: each basic emotion has an associated distinctive facial signal.
2. Distinctive physiology: each basic emotion is associated with a cluster of physiological properties, e.g. increased heart rate, perspiration, etc.
3. Automatic appraisal: each basic emotion includes an automatic appraisal of the environmental stimulus.
4. Distinctive universals in antecedent events: basic emotions developed as responses to species-constant problems related to survival. With time each basic emotion was associated with specific events e.g. sadness was associated with the loss of loved one.
5. Presence in other primates: each basic emotion can be found in and experienced by other primates.
6. Capable of quick onset.
7. Can be of brief duration; emotions are typically of brief duration.
8. Unbidden occurrence: each basic emotion cannot be stopped once activated.

In simple words according to BET a small set of specific emotions is hardwired in our brains, ready to be activated automatically when certain environmental challenges arise. Each program is associated with a distinctive facial signal, a dedicated neural circuit and a mandatory automatic response (Ekman 1999).

---

8 Ekman and Cordaro also include other characteristics such as the target of emotions and their destructive or constructive character. For a complete list of characteristics see Ekman and Cordaro (2011).
9 See Ekman (2003) for a detailed description of this characteristic.
10 I take neural circuit to mean a network or ensemble of neurons that process specific kinds of information, for example the myotatic (knee-jerk) spinal reflex is a common neural circuit where the stimulation of peripheral sensors initiates receptor potentials that trigger action potentials in specific neurons (see Gazaniga et al. 2014).
11 See the criticism raised by Nelson and Russell (2013), Russell (1994) and (2016) against the assumption that all emotions are associated with distinctive facial signals and mandatory responses.
connected with them, are evolutionary adaptations viewed as mechanisms developed to enhance survival rates. As such, every human should have them as they provide evolutionary advantages that contribute to our survival. For example we feel fear when encountering a bear in the woods so that we can move away from it and avoid being attacked. These emotions are activated automatically when certain environmental challenges arise such as dealing with dangers (fear), removing obstacles (anger), expelling noxious substances (disgust), suffering losses (sadness) and so forth (Scarantino 2015).

Researchers endorsing this theory rely heavily on experiments conducted by Tomkins in the 60s and Ekman in the 70s on the facial recognition of emotion (Tomkins 1962; Tomkins and McCarter 1964; Ekman 1972b). These experiments supposedly show that facial expressions relating to a small set of distinct emotions, namely fear, anger, sadness, joy, surprise and disgust, can be recognised by people universally. According to this view, we can prove the existence of universal emotions by showing that the recognition and production of the expressions associated with them are universal. Ekman’s research is supposed to prove the claim that facial expressions of some emotions are universally recognised by humans. According to this theory, then, the expressions and the emotions associated with them are universal.

Ekman argues that only the phenomena that fall into the category of basic emotions should be considered emotions (Ekman & Cordaro 2011). In other words, any phenomenon that does not share the characteristics associated with basic emotions is not an emotion for Ekman. Therefore his theory allows for the possibility that, if the category of emotion includes basic emotions exclusively, the category can be considered a natural kind.

An important detail of Ekman’s view is that basic emotions do not match the folk categories of emotion. Instead he claims that each of the basic emotions forms a family, for example under ‘anger’ we find instances of irritation, rage and fury (I use italics and quotation marks to refer to Ekman’s suggested basic emotions to distinguish them from folk emotion categories). Ekman argues that each category of basic emotion is a family of

12 Although there is disagreement amongst supporters of this theory as to which emotions qualify as basic, the most popular candidates included in the various suggested lists are the ones corresponding to the folk psychological emotion categories of Fear, Anger, Surprise, Sadness, Happiness and Disgust. However the list is currently being reviewed and other supporters of the theory of distinct emotions such as Izard provide a list of ten such emotions (Izard 1992).

13 The argument consisting of the premise that the recognition of facial expressions of emotion is universal and the conclusion that emotions are universal is not sound and I will reject it in chapter 4.
related states rather than a single state. What connects all the related states is the emotional theme. The emotional theme is what is hardwired, what is universal. To give an example, one of these basic emotions is ‘anger’ described by Ekman as follows:

‘Anger’: the response to interference with our pursuit of a goal we care about. ‘Anger’ can also be triggered by someone attempting to harm us (physically or psychologically) or someone we care about. In addition to removing the obstacle or stopping the harm, ‘anger’ often involves the wish to hurt the target. (Ekman & Cordaro 2011, 365)

According to Ekman, the following families of emotion are the basic emotions: ‘anger’, ‘fear’, ‘disgust’, ‘sadness’, ‘contempt’, ‘happiness’, and ‘surprise’. The basic emotion is identified with the emotional theme. For example ‘anger’ is the basic emotion, which can be manifested as irritation, or rage, etc. Irritation or rage under this account are subcategories of the basic emotion of ‘anger’ and they share the emotional theme which is supposed to be hardwired in humans. Ekman’s account predicted a correspondence between the emotional theme and the following:

- Dedicated neural circuit; this means that there is a distinct neural circuit corresponding to each basic emotion
- Distinct patterns of autonomic nervous system (ANS) activity, changes in physiology and behaviour for each basic emotion

Therefore, according to Ekman’s account, we would expect instances of rage and irritation to share a dedicated neural circuit and similar patterns of ANS activity, changes in physiology and behaviour. This is because a dedicated neural circuit and distinct patterns of changes are correlated with the emotional theme of ‘anger’.

In summary, according to Ekman’s version of BET, only certain specific phenomena qualify as emotions and could be considered natural kinds (or members of the same natural kind). As long as we only include phenomena that share the characteristics mentioned above, the overall category can be considered a natural kind.
1.2.1 Revised versions of Basic Emotion Theory

BET has been very influential in the philosophical debate on the natural kind status of emotion. For example Griffiths agrees with Ekman that some emotions can indeed be identified with basic emotions and that these emotions are the best candidates for natural kind status (Griffiths 1997, 245). However, Griffiths rejects Ekman’s view that emotion as a category should be identified with the category of basic emotions because if it did, ‘the extension of the emotion concept would be restricted to short-term, stereotyped responses, triggered by modular subsystems operating on a limited database’ (Griffiths 1997, 241). Instead, Griffiths argues that not all instances of emotion, or what he calls ‘the passivity’ phenomena, can be explained by the modularity of basic emotions (Griffiths 1997, 242). Griffiths holds that the vernacular concept of emotion includes at least two different subcategories of phenomena. These are the following (Griffiths 1997, 245-247):

1. Basic emotions (aka affect programs): Griffiths considers these as ‘short term, stereotyped responses, triggered by modular subsystems operating on a limited database’ such as fear, anger, joy etc. (1997, 245). These emotions are universal, modular and cognitively impenetrable in that their activation cannot be voluntarily controlled by cognition. They are characterised by a passivity that is explained in terms of modular information processing.

2. Higher cognitive emotions (aka irruptive motivations): States that are cognitively more sophisticated than basic emotions, e.g. guilt. Irruptive motivations require of the subject a certain mastery of complex concepts and they do not constitute direct responses to challenges presented by the immediate environment. They have no typical expressive alteration and they are not derived from more general goals by means-end reasoning. Some other examples would be loyalty or vengeance. Irruptive motivations allow people to adopt strategies that require them to be committed to act against their own interests should certain circumstances arise. For example a loyal friend involved in a life-threatening situation would still offer support to a friend despite experiencing fear, which would direct her to look for an escape route.

Griffiths actually suggests three different categories of emotional phenomena i.e. affect programs, irruptive motivations and disclaimed actions but he acknowledges that the third category should not be included in the same overall category with the former two (Griffiths 1997, 245).
According to Griffiths, the above categories have different phylogenies, different adaptive functions, different underlying neuroscientific properties and different roles in human psychology. Although Griffiths does not directly comment on the natural kind status of the category of higher cognitive emotions, he does state that the category of basic emotions is the best candidate for natural kind status. Answering the question of whether these two categories share any commonality, Griffiths accepts that both categories are states that interfere with the smooth unfolding of plans designed to secure our long-term goals (e.g. avoiding harm or loss). However, he claims that the psychological mechanisms underlying them are not of the same kind. Therefore he accepts a similarity of function between emotional categories but he also claims that there is no similarity in the underlying physical mechanisms that realise these functions. Whilst discussing what causes irruptive motivations he states:

Whatever psychological mechanism underlies the irruption of these clusters of desires into belief-desire causation, it is not the same mechanism that allows the affect programs to rapidly engage various effector systems without reference to consciously accessible beliefs and desires. (Griffiths 1997, 246)

Griffiths agrees with Ekman’s view that under each category of basic emotion, e.g. under the category of ‘anger’, we should only include instances which share the characteristics prescribed by BET as opposed to instances commonsense psychology groups together under anger. Any instances of anger that do not share these characteristics should not be included in the same category. As in everyday language we use the word ‘anger’ to refer both to instances of the basic emotion of anger and to the higher cognitive emotion of anger, Griffiths suggests that we should change the terms we use to reflect the difference in the referents. For example we could use the term ‘Affect program of Anger’ or ‘Basic Anger’ to refer to the instances of the basic emotion of ‘anger’ and some other term to refer to instances of the higher cognitive emotion of anger.

He concludes that we should eliminate the vernacular concept of emotion. He claims that it was introduced to refer to certain mental states that appeared to share a cluster of features (passivity, in the sense of being involuntarily instigated, being an important one among them). The causal mechanisms underlying the basic emotion of fear and the higher cognitive emotion of jealousy are too different to count these emotions as part of the same natural kind. Griffiths claims that this means that a) the term ‘emotion’ partially refers and b) that identifying the concept with either of the two categories seems unjustified
Griffiths concludes that the concept of emotion is not a useful concept for affective science. He states:

My central conclusion is that the general concept of emotion is unlikely to be a useful concept in psychological theory.....There is no rich collection of generalisations about this range of phenomena that distinguishes them from other psychological phenomena. (Griffiths 1997, 14)

Griffiths therefore claims that there is no object of scientific knowledge which corresponds to the term ‘emotion’ (Griffiths 1997, 16) and, as such, we should replace the category with two different categories in a future psychology of emotion (Griffiths 1997, 229).

In the passages quoted above Griffiths assumes that only categories that can be considered natural kinds should be objects of scientific investigation. This assumption is questionable; however, in this thesis I will accept this assumption and focus on the question of whether emotion is a natural kind rather than the different question of whether categories need to be natural kinds in order to have explanatory and predictive power in scientific and non-scientific theory.

Therefore for Griffiths the category of emotion is not a natural kind and the vernacular concept of emotion serves no explanatory purpose in science. Certain instances of individual emotions (‘finer’ grained emotions that share the characteristics stated by Ekman), can be considered individual natural kinds and the category of basic emotions could also be considered a natural kind (Scarantino 2015). According to this view, these ‘finer grained’ emotions e.g. ‘the basic emotion of Fear’ refer to phenomena sharing identical underlying (neurobiological) causal mechanisms and, as such, these emotions are neurobiological kinds.

Hutto (2018) recently suggested an alternative version of BET according to which we should interpret the distinction between basic emotions and non-basic emotions as one

---

15 Although Griffiths makes both claims one could argue that these are contradictory; if emotion partially refers then identifying it with one of the two categories to which it refers is actually justified. I discuss and reject the claim of partial reference in chapter 3.3.1.

16 Goldie for example argues that even if emotion is not a natural kind this does not entail that the category lacks any explanatory or predictive power (see Goldie 2000, 103). The answer one gives to this question depends on how one defines ‘natural kinds’. As will become evident in chapter 2, the account I endorse identifies natural kinds as categories which necessarily show an aptness for successful induction and explanation and therefore natural kinds by definition are categories with explanatory and predictive power.
that tracks extra ingredients. The category of non-basic emotions simply has an added ingredient that the category of basic emotions lacks; the extra ingredient is ‘content’. According to Hutto, it is not the case that basic emotions are hardwired, or ‘brain bound’ (Hutto 2018, 14). Instead ‘basic emotions can be understood in adaptive non-reductive terms, which nevertheless lack the contentful features of non-basic emotions’ (Hutto 2018, 14). It is not clear that this view can be considered a continuation of BET as it rejects the claim that emotions are hardwired (biologically basic). As numerous quotations from Ekman and Scarantino show in this chapter, one of the commitments of BET is that basic emotions are hardwired, automated programs. It is not clear what alternative sense of ‘basicness’ Hutto’s view supports and I will not assess it in this thesis as I focus on the biological sense adopted by Ekman, Scarantino and Griffiths.

1.3 View B: The category of emotion is a natural kind and some categories of individual emotions are natural kinds

Griffiths’ influential view suggests that there is no unity of emotion. Instead the category of emotion is fragmented into at least two different subcategories. Several philosophers have attempted replies to the disunity claim so as to preserve the natural kind status of the category (Prinz 2004; Charland 2002; Roberts 2003; Deonna & Teroni 2012). The view shared by these philosophers is that emotion as a category is a natural kind. However they disagree with one another on the basis of this claim because they disagree on the nature of emotion. Prinz for example argues that emotion is a natural kind because all phenomena included under this term are perceptions (he calls them ‘embodied appraisals’) (Prinz 2004, 102), whilst Deonna and Teroni (2012, 25) reach the same conclusion by arguing that all emotions are attitudes. Both theories are considered compatible with BET and they integrate this theory in their hypotheses for the natural kind status of emotion.

Prinz claims that the category of emotion is a natural kind because all emotions are basic emotions or blends of basic emotions (Prinz 2004, 97) and they are states of the same type (Prinz 2004, 102). He therefore argues that we can preserve the unity of emotion if it can be shown that:

1. All emotions are basic emotions or
2. All emotions are blends between basic emotions and non-basic emotions
Prinz’s claim that emotions are perceptions is based on the theory of William James who claims that ‘the bodily changes follow directly the perception of the exciting fact, and that our feeling of the same changes as they occur is the emotion’ (James 1884, 19). According to James’ theory, emotions are feelings we have of changes occurring in our body following the perception of a stimulus. To give a few examples, according to James we are angry because we tremble rather than trembling because we are angry, and we are afraid because we cry rather than crying because we are afraid. This suggestion goes against folk psychology interpretations of emotion where we usually think that bodily changes follow the emotion and are not followed by it.

Prinz argues that all emotions share the following characteristics (2004, 102):

- All emotions are perceptions and if so, all emotion is perception.
- All emotions are associated with expressive behaviour
- All are associated with autonomic responses (that is responses of the autonomic nervous system (ANS))
- All seem to involve related structures in the central nervous system
- All are eruptive
- All are associated with action tendencies (e.g. to run, to freeze, to attack, etc.)
- All are motivating
- All engage attention
- All are valent (they are evaluated as positive or negative)

Prinz accepts Griffiths’ subcategories of affect programs (basic emotions) and irruptive motivations (non-basic emotions) but contra Griffiths he claims that these two subcategories share the following:

1. Affect programs and irruptive motivations demand the same explanatory resources.
2. Affect programs and irruptive motivations are states of the same type in that they are caused by the same thing (e.g. same type of underlying mechanism).

Prinz therefore claims that all emotions are members of the same overall category because they share the same type of causal mechanism and demand the same explanatory resources. Therefore, although Prinz accepts Griffiths’ subcategories of emotion, he
disagrees with him because he claims that all emotions share the same type of causal mechanism. His theory also allows that individual categories of emotions can be considered natural kinds.

Deonna and Teroni integrate BET in their theory in a similar way by claiming that the category of emotion is likely to be a natural kind, if all phenomena we consider to be emotions turn out to be basic emotions. They claim that we may discover that there are more basic emotions that we thought there were:

One cannot exclude the possibility, for example that instances of an emotion dependent on a certain complex judgement share similar underlying causal mechanisms, and therefore qualify as affect programs insofar as they implicate all the physiological and neural changes characteristic of them…..While some have appealed to this distinction (between basic and non-basic emotions) in support of the thesis that this category [that is, emotion] exhibits no interesting unity, we have argued that there are reasons to resist this view. (Deonna & Teroni 2012, 25)

Deonna and Teroni agree with Prinz’s position on a shared similar causal mechanism. However, they disagree with Prinz’s claim that all emotions are perceptions. Instead they claim that all emotions are attitudes. They state:

An emotion is an attitude towards an object, an attitude which it is appropriate to have when the latter exemplifies a given evaluative property. (Deonna & Teroni 2012, 76)

What is important in this theory is not the evaluative feature of the object but the evaluative attitude of the subject. This is why it makes sense to say that the same stimulus may cause different emotions, e.g. what frightens Julianne is what Mary is amused by – a dog, for instance (Deonna & Teroni 2012, 77). They explain that:

It is for instance because Julianne takes the attitude of fear towards the dog that its dangerousness features in the correctness conditions of her mental state. (Deonna & Teroni 2012, 77)

In the example given above, the dog is amusing rather than dangerous and as such, Julianne’s fear is not an appropriate attitude to have towards it. Deonna and Teroni argue that emotion can be considered a natural kind if all emotions are evaluative attitudes caused by evaluative properties of objects. Their theory seems to suggest that emotion could be a
natural kind and that categories of individual emotions can also be considered natural kinds. For the purposes of this thesis what is important is that they integrate BET into their theory by suggesting that more emotions may turn out to be basic emotions.

1.4 View C: Emotion is not a natural kind. Individual emotions are not natural kinds.

According to a third view, emotion is not a natural kind and individual emotions, the ones corresponding to folk categories or the ones corresponding to basic emotions, are not natural kinds either. This sceptical view is associated with psychological constructionism, namely the view that emotions are constructed from the combination of other psychological processes such as attention and memory, which are not unique to emotion (Barrett 2006b; Russell 2015).

Constructionists make this claim on the basis of different theories about the nature of emotion. Averill for example views emotion as a transitory social role, that is, a prescribed social role to be followed by a person in a given situation (Averill 1980, 308). Modern constructionists, such as Russell (2015) and Barrett (2006b), claim that the agent ‘constructs’ emotional meaning through the categorisation of primitive core affect (viewed as the combination of valence (positive or negative) and physiological arousal). Barret and Russell claim that emotion is not a natural kind and that the so-called basic emotions (individual basic emotions or the category that includes them) are not natural kinds (Barrett & Russell 2015). Barrett suggests instead that valence or arousal could be more likely candidates for natural kind status.18

---

17 Barret and Russell disagree on whether the categorisation is part of the emotion or a consequence of the emotion. Russell opts for the latter, whereas Barrett includes the categorisation process within the emotion. This difference is not important for the main issue under discussion here because they both agree that emotion is not a natural kind and that basic emotions are not natural kinds.

18 Whether valence or arousal can be considered natural kinds is not the focus of this thesis. However with respect to valence see Colombetti (2005) who argues that ‘valence’ is a complex concept which should not be reduced to one simplified scientific definition as this distorts phenomenological experience.
1.5 Assumptions shared by main views

It is interesting to note that researchers who may have similar views on the nature of emotion, may disagree on the issue of the natural kind status of the category. To give an example, both Prinz and Barrett claim that their theories are based on James’ perceptual theory of emotion. However, they disagree on the following:

1) Prinz accepts BET but Barrett rejects it.
2) Prinz claims that the category of emotion is a natural kind whilst Barrett suggests the opposite.
3) Prinz’s theory allows for the category of basic emotions to be considered a natural kind but Barrett rejects this view.

On the other hand, researchers who may disagree on the nature of emotion may agree on the natural kind status of the category. For example, Prinz takes emotion to be a perception whilst Deonna and Teroni take emotion to be an attitude. Both suggest that emotion could be considered a natural kind and both accept BET theory and integrate it in their accounts. Therefore similar theories held on the nature of emotion do not necessarily correspond to similar theories held on the natural kind status of the category, while, on the other hand, different views on the nature of emotion sometimes agree on the natural kind status of the category.

Secondly, all the positive views mentioned above, that is views A and B, seem to accept the claim that emotion or some of its subcategories are neurobiological kinds. In the same way the negative view, view C, claims that emotion and individual emotions are not neurobiological kinds. According to view C, emotion is not a category whose members share either a distinct property cluster of neuro-physiological properties or a biological signature (an underlying distinct neurobiological mechanism). Therefore the focus of this debate seems to fall on the question of whether emotion is (or could be) a neurobiological kind; and this focus is shared across the different views in this debate. If emotion was a neurobiological kind then we would expect it to have a distinct cluster of neurobiological properties (or an underlying neurobiological causal mechanism) which differentiate it from other neurobiological kinds such as cognition. In a similar way the chemical category of metal for example has a distinct cluster of chemical properties e.g. high electrical and
thermal conductivity, which distinguish it from other chemical categories e.g. the category of noble gas.

An additional view held by all sides in this debate is the separate claim that if emotion is not a neurobiological kind, it is not a natural kind; instead the negative views hold that it is a social kind (Barrett 2006b; Griffiths 1997) and, as such, it is not a natural kind. This depends on the separate metaphysical assumption that social kinds cannot be considered natural kinds. In this thesis I argue against both of these very popular assumptions held in affective science:

1) Emotion or some of its subcategories, namely individual emotions, are neurobiological kinds: this view takes emotion to be identified with and explained by neurobiological properties.
2) Social kinds cannot be considered natural kinds.

In Part 1 of this thesis I look into and reject the view that it can be established a priori that social kinds are not natural kinds. I argue that whether or not they are natural kinds is a question to be determined a posteriori. In Part 2 I look into and reject the view that emotion (and emotions) is a neurobiological kind and that emotion can be reduced to and identified with neurobiological properties. Instead, I argue that emotion and emotions are social kinds. Although emotional phenomena include neurobiological properties, they are not exhausted by them and should not be reduced to them. Emotion is an interactive phenomenon with a range of properties in social psychology, sociobiology and social ecology; its function is to enable a successful social life. My view assumes an externalist view of mental states in general and emotionality in particular. Some of the properties which combine to create emotion are importantly external in that the person who experiences them needs to interact with the environment so that they are created and developed. In other words I claim that the human emotional apparatus is not fully developed at birth. It needs to be developed via interaction with the environment so that we experience emotion. My view borrows from Greenwood’s (2015) analysis of the ontogenesis of emotion, but my focus falls on classifying the category of emotion at a higher level than the purely neurobiological.

Although the main focus of this thesis is to provide an answer to the metaphysical question of whether the category of emotion is a natural kind, this metaphysical issue has very important epistemological consequences for affective science. In the next section I
show that this debate impacts affective science by guiding experimental design and
determining the interpretation of empirical data. Theories held implicitly or explicitly by
researchers on whether emotion is a neurobiological natural kind, heavily influence what
experiments scientists design and also how they interpret experimental data.

1.6 Natural kind status of emotion and empirical science

Barrett claims that the view that emotion is a natural kind is an assumption widely accepted
in psychology (2006b). However, she argues that this assumption is not justified by the
available empirical evidence. She states:

   I suggest that the natural-kind view of emotion may be the result of an error of
   arbitrary aggregation. That is, our perceptual processes lead us to aggregate
   emotional processing into categories that do not necessarily reveal the causal
   structure of the emotional processing. I suggest that, as a result, the natural-kind
   view has outlived its scientific value and now presents a major obstacle to
   understanding what emotions are and how they work. (Barrett 2006b, 29)

Even if someone does not agree with Barrett’s claims on the natural kind status of emotion,
she is right to highlight the fact that, implicit or explicit theories on the natural kind status,
have an impact on the design of experiments on emotion and also, on the interpretation
of experimental data. Researchers who argue that emotion is a neurobiological natural kind
seem to accept the following assumption:

   Emotion as a category is a natural kind because all members share the same
   cluster of properties in physiology, behaviour and phenomenology which are
   caused by the same underlying mechanism (e.g. Charland 2002; Prinz 2004).

In a similar way, researchers who argue that categories of individual emotions can be
considered natural kinds, accept one or both of the following assumptions:

1) Categories of individual emotions, such as fear, can be considered natural kinds
because all instances of certain specific emotions share an underlying causal
mechanism viewed as a distinct neural circuit dedicated to each emotion (e.g.
Damasio 1994; Ekman 1972b; Izard 1992; Panksepp & Biven 2012; Matsumoto
& Willingham 2009)
2) Categories of individual emotions, such as fear, can be considered natural kinds because all instances of this emotion share the same properties in physiology, behaviour and phenomenology (e.g. Scherer 1994).

Barrett argues that on the basis of the assumptions above, that is the existence of a property cluster or an underlying causal mechanism, scientists tend to assume that any single response (e.g. facial movements alone) can be taken as evidence that a particular kind of emotion has occurred (Barrett 2006b, 32). Therefore one response can function as a proxy for the others. She argues that:

As a result, scientists feel free to state that their findings are about, for example, anger even if they measured only a peripheral nervous system response, facial movements or a subjective feeling state. If one behaviour, (say a voluntary action) is present but another (say, facial expression) is not, then it is assumed that the tendency was there, but that some other mechanism (e.g. a display rule) interfered with its expression. (Barrett 2006b, 32)

If emotions are identified with automatic response programs sharing a cluster of similar properties on the basis of an underlying causal mechanism, then we would expect the existence of one of the properties from the cluster to necessarily imply the existence of more properties from the same cluster.\(^{19}\) So if anger, for example is associated with specific properties distinctive of the emotion of anger (let’s say properties E, F, G), if we observe the existence of property E, we could safely infer that properties F and G are also present, and that the emotion experienced is therefore an instance of anger. As Barrett describes above, if empirical observation does not confirm this hypothesis, researchers postulate that a separate additional mechanism obstructs emotional expression, so that properties F and G are present but somehow masked, instead of rejecting the original hypothesis.\(^{20}\)

An alternative response to the apparent absence of F and G would be to doubt the truth of the original hypothesis, which infers the presence of properties F and G from the observation of property E. Therefore, theories held on the nature of emotion influence both the design of experiments and the interpretation of experimental data.

---

\(^{19}\) In chapter 2 I analyse such views in detail when discussing Boyd’s (1991) property cluster account where a cluster or similar properties is maintained by similar underlying causal mechanisms.

\(^{20}\) Ekman’s theory of ‘display rules’ is an example of a theory suggested as a means to compensate for the absence of predicted properties. See Ekman (1972b) for more detail on ‘display rules’.
In what follows I discuss a well-known experiment (Schachter & Singer 1962) and a recent meta-analysis (Lench et al. 2011) to show how researchers holding different theories on the natural kind status of emotion interpret the same empirical results in different ways to reach opposite conclusions.

1.6.1 Interpreting experimental data and designing experiments on emotion

Schachter and Singer (1962) conducted a hugely influential experiment which, nonetheless, has not been replicated since. In the experiment subjects were injected with adrenaline but were told that it was a special vitamin which may improve eyesight. They were then asked to sit in a waiting room for a short period of time before going through an eye exam. Adrenaline causes increased autonomic responses such as tachycardia (heart racing), heavy breathing and perspiration. Some subjects were told to expect side effects and some were not. Some of the subjects were told to fill out a questionnaire while they were waiting for the eye exam. The questions on the questionnaire were progressively more insulting, culminating in questions such as ‘With how many men (other than your father) has your mother had extramarital relationships?’ with the answer options being a) 4 and under, b) 5-9 and c) 10 and over. A fake participant who has been placed in the waiting room expresses increased aggravation and finally storms out, tearing up the questionnaire. Other subjects have a very different experience in the waiting room where they do not receive offensive questionnaires but are accompanied by a fake participant who is acting silly (making paper airplanes, playing with hula hoops etc.).

The results of the experiment show that participants who were informed about the side effects exhibit little emotional response (in behaviour or reported experience) in either condition (questionnaire or no questionnaire). Participants who were not informed about the side effects were observed to express outrage in the questionnaire condition. Uninformed participants who did not receive questionnaires exhibit signs of euphoria. The data of the study was based on observation of behaviour and on the reported experience of participants. The experimenters observed the behaviour exhibited by the participants to reach conclusions about their emotional state. So participants who exhibited behaviour typical of anger were considered to feel angry, etc. This means that the experimenters assumed that participants who exhibited a specific emotional behaviour or reported an emotional experience, e.g. angry or happy, were actually in the same physiological state.
According to Schachter and Singer, the results show that although the same bodily states were induced by the injection of adrenaline, they were interpreted differently by the participants who experienced either anger or euphoria. How these bodily states were labelled was determined by context (such as the questionnaire, other participants) and background knowledge (information about side effects). According to Schachter and Singer, the experiment shows that interpreting these bodily states as one emotion or another is a cognitive process.

Schachter and Singer used this experiment as support towards their ‘two factor’ theory of emotion, which is based on James’ theory. According to their theory, a stimulus produces physiological arousal (factor one). Subsequently via a cognitive process (factor two) we explain/interpret the arousal in relation to the stimulus and this produces an emotional experience. To give an example, seeing a snake five feet away from you elicits the same physiological arousal as winning £100K in the lottery. It is only after this arousal has been related to the snake or to the lottery win, by interpreting the connection between the arousal and the stimulus, that the feeling is interpreted as one of fear or joy. According to this view, increased arousal is perceived by the individual as requiring an explanation and so a cognitive analysis of contextual features results in a differentiated feeling state (Sander & Scherer 2009).

Researchers who hold different assumptions about the natural kind status of emotion, such as Prinz, claim that several assumptions accepted by Schachter and Singer can be challenged (Prinz 2004, 70). Firstly, if the participants actually went through emotional experiences, there is no reason to suppose that these would not modify in some way their physiological states induced by the injection of adrenaline. For the physiological states of the participants to remain the same, as Schachter and Singer claim that they do, one needs to presuppose that the physiological state induced by the injection of adrenaline is not modified in any way when we go through anger or joy. As Prinz points out, this is an assumption that is far from secure. Once the emotional response occurs, either of anger or of joy, ‘there is no reason to assume that the physiology in the amusing and aggravating situations will remain the same’ (Prinz 2004, 71). As the physiology of the participants was not monitored there is no way for us to know which hypothesis was right in that case. Therefore it remains possible that, as BET predicts, the emotion of anger would necessarily be identified with the activation of an automatic programme of responses in physiology. If so, the physiology of the participants, as soon as anger was activated, would necessarily change from the physiology they had before they experienced anger.
Researchers who endorse BET predict that the physiology observed between a participant undergoing an instance of anger and a participant undergoing an instance of joy will necessarily be different (Ekman & Cordaro 2011). On the other hand, researchers who endorse constructionist or componential theories predict that the underlying physiology of anger and joy will be similar and not necessarily different (Schachter & Singer 1962; Barrett & Russell 2015). In the absence of the relevant physiological data it seems to be the case that these predictions lead to different interpretations of the same experimental results.

Secondly, although participants exhibited different output emotional behaviour, namely of anger or joy, when asked to report on their emotional experience, participants in both conditions (questionnaire, no questionnaire) reported that they were happy. The same feeling was reported by participants who were informed about the side effects. However, we would expect participants who were injected with adrenalin and then went through the questionnaire condition to report feelings of anger. Schachter and Singer tried to answer this criticism by suggesting that the participants were trying to please the experimenter to receive extra points on their final exam (and in effect lie about their feelings when they reported their experience) and did not report feelings of anger. They state:

As we subsequently discovered… the subjects, who had volunteered for the experiment for extra points on their final exam, simply refused to endanger these points by publicly blowing up, admitting their irritation to the experimenter's face or spoiling the questionnaire. Though as the reader will see, the subjects were quite willing to manifest anger when they were alone with the stooge, they hesitated to do so on material (self-ratings of mood and questionnaire) that the experimenter might see and only after the purposes of the experiment had been revealed were many of these subjects willing to admit to the experimenter that they had been irked or irritated. (Schachter & Singer 1962, 391)

So Schachter and Singer claim that self-reports were influenced by a desire to please the experimenters and suggest that perhaps they should not be taken at face value. However one can claim the same about the observed behaviour of the participants and claim that the participants were influenced by a desire to please the stooges and faked their
emotional behaviour. If so, the observed emotional behaviour should not be taken at face value for the same reason that self-reports should not be taken at face value.

Another more recent example of different interpretations of experimental results relates to a meta-analysis of 687 studies conducted by Lench et al. in 2011. Lench et al. analysed the studies on emotion elicitation by the use of film, pictures, music, imagination and other methods to test the hypothesis that each ‘basic emotion’:

- elicits changes in cognition (e.g. narrowing of attention on a tiger in the distance), judgment (e.g. the risk perceived in the environment), experience (e.g. the recognition that one is afraid), behaviour (e.g. a tendency to run away), and physiology (e.g. increased heart rate and respiration). (Lench et al. 2011, 835)

Lench et al. acknowledge that their review included only studies that attempted to elicit discrete emotions and that it could not be informative about the effect sizes associated with elicitations of general positive or negative states (Lench et al. 2011, 850). When discussing the results of their review they state:

In the present investigation, the effect sizes associated with comparisons among all emotions were consistently greater than zero only for self-reported experiences. For measures other than self-report, the effect sizes that compared anxiety and anger, sadness and anger, and sadness and anxiety were small. (Lench et al. 2011, 850)

Lench et al. accept that the result above suggests that differences between discrete negative emotions may be the result of beliefs about emotions (2011, 850). They also accept that there was no evidence that anxiety influenced outcomes other than self-report in ways that differed from other negative emotions. However, according to BET we would expect to find consistent and clear differences between basic emotions in the physiology, behaviour and reported experience of the participants. Lench et al. acknowledge that:

---

21 Following methodological progress in the last 20 years, we are now able to observe the underlying physiology of participants through the use of electromyographic (EEG) data or functional magnetic brain imaging (fMRI) data along with collecting experience reports and observing the behaviour of participants. Scientists endorsing BET design experiments attempting to capture regularities in physiology (via EEG and fMRI) between instances of distinct emotion categories. To be able to create correlations between clusters of physiological changes and distinct emotions they typically use reports provided by participants. Therefore the use of the reported experience of participants is important in order to create the types of systematic correlations postulated by BET.
If differences were due to beliefs about emotions, differences among discrete emotions would be expected to be most evident in self-reports of emotional experience that are at least partially based on general beliefs about emotional experience and least evident for other outcomes, such as behavioural and physiological reactions. (Lench et al 2011, 835)

This is precisely what their results show; the effect size for ‘Self-report experiential’ outcome is quite large, namely 0.83, whereas the effect size for behavioural and physiological outcomes is 0.31 for each outcome (Lench et al. 2011, Table 1, 843).

However, and in spite of the claims made in the quotes above, Lench et al. conclude that their meta-analysis provides support for BET because it shows that there are differences between discrete emotions (positive/negative), and between discrete negative emotions (anxiety/anger); and that there were correlated changes in the behaviour, experience and physiology of participants.

By contrast, Lindquist et al. (2013) analyse the data accumulated by Lench et al. (2011) to reach a different conclusion. They claim that the findings by Lench et al. do not ‘address whether specific emotions have diagnostic clusters or outputs, as is hypothesized by the discrete emotion models cited in their paper’ (Lindquist et al. 2013, 5). Instead they claim that:

The only consistently significant differences (with moderate effect sizes) across pairwise comparisons of negative emotion inductions in Lench et al.’s (2011) meta-analysis were observed for self-reported emotional experience (effect sizes range from .24 -1.61, p’s <.001; see Table 3.) The fact that participants report differences in experience across different emotion inductions, but do not show consistent and specific differences in physiological responding and behaviour, is more consistent with the alternative hypothesis that emotions are constructions of the human mind – complex perceptions where physiological responses are made meaningful in context. (Lindquist et al. 2013, 5)

This is another example of a case where different theories held on the natural kind status of emotion influence the importance researchers attach to certain experimental data. Lench et al. seem to predict that emotions have clear and distinct boundaries, they are neurobiological categories with observable clusters of properties:
To the extent that emotions represent discrete constructs, each emotion should differ from other emotions (and neutral groups) on measures of cognition, judgment, experience, behaviour and physiology and the differences should co-occur across these systems. (Lench et al. 2011, 835)

On the other hand, Lindquist et al. (2012) take emotions to be categories in folk psychology. According to Lindquist et al. (2012), emotions are not natural categories in neurobiology with clear-cut boundaries and distinct underlying physiological signatures. They state:

In a psychological construction framework, emotions are not special mental states with diagnostic patterns of output. Instead, the hypothesis is that an emotion word names a commonsense category that corresponds to a wide range of mental events that vary in physiology, behaviour, cognition and experience….Not all instances of an emotion referred to by the same word (e.g. anger) look alike, feel alike, or have the same neurophysiological signature. (Lindquist et al. 2013, 6)

Given the different views held on the natural kind status of emotion, Lench et al. (2011) opt for interpreting the evidence as supporting BET, whereas Lindquist et al. (2013) interpret the evidence as supporting constructionist models.

As the discussion of the case examples shows, what at first appears as a metaphysical and epistemological exercise in one aspect of the nature of emotion, turns out to have implications for experimental psychology. The same considerations should apply to other scientific disciplines dealing with emotion such as clinical psychology, psychiatry or AI. To give an example from psychiatry, if emotion is a neurobiological kind and its members share a cluster of neurobiological properties or the same underlying neurobiological mechanism, this will dictate how emotional disorders are treated by clinical psychologists and psychiatrists; if emotion qua neurobiological kind can be reduced to neurobiological or neurochemical properties, these properties can be manipulated by chemical means such as drugs (in other words a chemical cause calls for a chemical cure). If it cannot, although their use may ameliorate some of the symptoms of emotional disorders it may not remove the cause.

This thesis agrees with constructionist or componential accounts of emotion which predict that we will not find a shared underlying neurobiological causal mechanism
responsible for emotion or emotions. However I disagree with the constructionist assumption that, if emotion is not a neurobiological kind, this entails that it is necessarily not a natural kind. If the hypothesis advanced by this thesis is correct, we need to approach the science of emotion using an alternative account of natural kinds.

1.7 Summing up

In this section I introduced three different views on the natural kind status of emotion. Although these are not the only possible views on the natural kind status of emotion they are the most dominant views in affective science. According to the first one (View A) some categories of emotion are natural kinds and the category which exclusively includes these categories could be considered a natural kind. According to the second one (View B) some categories of emotions are natural kinds and the category of emotion which includes these and combinations of these emotions is a natural kind. According to the third one (View C) the category of emotion and individual categories of emotion are not natural kinds. I argued that all three views share two important assumptions on the nature of emotion and the metaphysics of natural kinds. These are the following:

- Emotion or its subcategories are neurobiological kinds; they can be explained by or be identified with neurobiological properties
- Social kinds cannot be considered natural kinds

In the second part of this section I argued that implicit or explicit views held on the natural kind status of emotion influence the design of experiments in experimental psychology. Researchers who accept that the members of some categories of emotion share identical clusters of properties, on the basis of sharing similar underlying causal mechanisms, design experiments with a view to discovering these clusters and underlying causal mechanisms. In addition, I claimed that depending on the view held by researchers on the natural kind status of emotion, or of particular emotions, they focus on specific experimental data when interpreting results; researchers endorsing BET focus on the similarity observed in the neurophysiology, behaviour and reported experience between similar instances of individual emotions. On the other hand, researchers who endorse constructionist or componential views on the nature of emotion, focus on the variety shown in the neurophysiology, behaviour and reported experience between similar instances of individual emotions. I suggested that the issue of the natural kind status of emotion is not simply an issue in metaphysics with no practical application in empirical science. Therefore, if one shows that the assumptions listed above – that emotion is (and
particular emotions are) a neurobiological kind and that social kinds cannot be considered natural kinds – are unwarranted, a further view on the natural kind status of emotion becomes possible: emotion is a social natural kind.

In what follows I divide this thesis in two parts. Part A deals with the metaphysics relating to natural kinds (chapter 2) and social kinds (chapter 3). Part B focuses on the category of emotion and contains two claims. In chapter 4, I claim that emotion is not a neurobiological kind, rejecting views A and B discussed in this introductory chapter. In chapter 5, I claim that emotion is a social kind, thus agreeing partly with view C. However, I disagree with View C on the natural kind status of emotion as I conclude that according to the account of natural kinds endorsed in chapter 2, emotion can be considered a natural kind.
Part A - Natural kinds and social construction

Chapter 2 – Towards a plausible account of natural kinds

In this chapter I discuss two different views on natural kinds, the ‘metaphysical view’ and the ‘epistemological view’ to argue that a specific version of the epistemological view provides a plausible account of natural kinds and is better equipped to explain the use of natural kinds in scientific and non-scientific disciplines. In the first part I show why natural kinds are considered useful categories by showing their aptness for induction and explanation. I then assess which of the two rival views provides a better explanation of this aptness for successful induction and explanation to conclude that the epistemological view provides a better account of natural kinds. Finally, I endorse a recent property cluster account suggested by Slater (2015) and argue that it has advantages over other property cluster accounts suggested by Boyd (1991) or Khalidi (2013).

2.1 The metaphysical view vs the epistemological view

According to what I call the ‘metaphysical view’, natural kinds are entities possessing essential properties; the possession of these properties is a necessary condition for membership into the kind (Ellis 2001, 19; Bird 2009). According to this view, these kinds are independent of humans in the sense that they exist in nature independently of us. These kinds are metaphysically fundamental in the sense that they constitute the ‘joints’ in nature; they are fundamental compared to artefact objects created by us such as chairs or cars. On the other side of the debate, according to what I call the ‘epistemological view’, natural kinds are categories which show an aptness for successful induction and explanation. They are epistemically ‘fundamental’ categories because they facilitate our knowledge and understanding of the phenomena under investigation. Their status derives from the epistemic role they play regardless of whether they correspond to any fundamental entities dependent wholly on nature. Instead, according to the epistemological view, natural kinds are epistemological tools which advance our knowledge and understanding of various phenomena (Boyd 1991; Dupré 1993; Khalidi 2013). According to the epistemological view natural kinds partly depend on the theoretical domains they are embedded in rather than matching already existing divisions in nature.
Although a consensus on the definition of ‘natural kinds’ is currently missing, this term is used extensively in philosophy and in several other disciplines such as biology, psychology and medicine. Hacking (2007, 205) argues that we need to reach a consensus on the meaning of the term to avoid conceptual confusion and achieve progress. In this chapter I claim that the epistemological view is more plausible than the metaphysical one and that a recent account suggested by Slater (2015) provides a good basis for an epistemological account of natural kinds.

In the first part I discuss certain features associated with natural kinds to focus on the relation between natural kinds and inductive inference. Natural kinds are considered important for scientific disciplines because they supposedly provide a justification for scientific inferences and practices (Bird & Tobin 2018, §1). I then focus on the two different accounts of natural kinds described above. I argue that a combination of the epistemological view with pluralism, the thesis that there are many legitimate ways ‘to carve nature at the joints’ (Dupré 1993), provides a plausible account of natural kinds. According to this account, natural kinds are kinds which show an aptness for successful induction and explanation in accordance with the practices and methodologies of the theoretical domains those kinds are embedded in. These kinds are epistemically fundamental as they facilitate the inductive and explanatory inferences required by the relevant theoretical domains. According to the epistemological view, kinds can be found everywhere in nature, including the social domain, and they can include artefact kinds such as chairs or cars. I conclude that although the epistemological account, as recently modified by Slater (2015), seems to be overly permissive in allowing artefact kinds, psychological or social kinds to count as natural kinds, nonetheless, it still preserves the usefulness of the concept of natural kinds.

2.2 Natural kinds, induction and explanation

Both sides in this debate use the categories of chemical elements as prime examples of natural kinds. Chemical elements are divided on the periodic table (see figure 1 below) on the basis of atomic number (the number of protons in the nucleus of the atom). So for example hydrogen, $H$ at the top left has one proton in the nucleus of the atom whereas helium, $He$ at the top right has 2 protons in the atom nucleus. Another criterion that differentiates atoms from one another is atomic mass (the sum of protons and neutrons

---

22 I use italics to refer to the kind rather than the term throughout this chapter.
in the atom nucleus). However, this difference is not reflected on the periodic table which specifically tracks atomic number. As such, elements with the same atomic number but different atomic mass (called isotopes) e.g. Carbon 12 and Carbon 13, are included in the same place on the periodic table; both atoms have the same number of protons (6) in the nucleus but they are not identical, in the sense of having identical microstructure, because they have different numbers of neutrons in the nucleus e.g. Carbon 12 has 6 and Carbon 13 has 7.

Figure 1: The periodic table

In this example, atoms which do not have identical microstructure e.g. Carbon 12 and Carbon 13, are nevertheless included in the same category of carbon. The reason for this is that having the same atomic number guarantees that all isotopes of Carbon will have specific chemical properties which chemists are interested in investigating. We can create successful generalisation and induction for isotopes of elements e.g. isotopes of carbon or silver on the basis of the fact that elements possessing the same atomic number, have identical properties. Examples of such induction and explanation based on observation of individual instances can be found below:

- Silver Ag has a boiling point of 2162°C (from observing several examples of silver).
• Silver is less malleable than gold (from observations made on examples of silver and comparisons with examples of gold).

• Silver is used for cutlery because of its antibacterial properties (from observing that silver irreversibly damages key enzyme systems in the cell membranes of pathogens but its toxicity to human cells is very low).

Although there is no consensus as to what determines natural kind status, several conditions are associated with natural kinds. According to Bird and Tobin, the characteristics listed below are traditionally associated with natural kinds (Bird & Tobin 2018, §1):

1. **Members of a natural kind should have some (natural) properties in common**: It has been suggested that all members of a kind should necessarily share some intrinsic properties.\(^{23}\) In the example of silver mentioned above, such properties could include atomic number, boiling point, melting point etc. It could be the case that one of the properties, such as atomic number is causally responsible for the rest of the properties or it can explain the existence of all other properties.

2. **Natural kinds should permit successful inductive inferences**: We should be able to make successful generalisations that apply to all members of a kind on the basis of making observations on single members of the kind.

3. **Natural kinds should participate in laws of nature**: According to Ellis, the laws of nature describe the essential properties of the natural kinds (Ellis 2002, 82). To say that all members of a class share the same property e.g. ‘All As are Bs’ is to reveal a universal proposition (law of nature); a proposition which refers to all members of the class and it is true of all of them (Ellis 2002, 89).

4. **Members of a natural kind should form a kind** (Bird & Tobin 2018, §1.1.1): According to this condition, the natural kind formed should correspond to a natural grouping in nature and not to a gerrymandered kind. Bird and Tobin acknowledge that it is not clear what this condition requires however they use the example of ‘white things’ or ‘negatively charged things’ as examples of kinds that would not qualify under this condition (Bird & Tobin 2018, §1.1).

---

\(^{23}\) Bird notes that often philosophers take ‘natural properties’ to mean ‘intrinsic properties’; however there are philosophers who disagree (Bird & Tobin 2018, §1.1).
5. **Natural kinds should form a hierarchy:** In cases where two kinds overlap by sharing some but not all properties with one another, one is a subkind of the other (Ellis 2002, 82). For example, the natural kind *metals* and the natural kind *zinc* satisfy this condition. Members of the natural kind *metal* share some properties with one another but they do not share all their properties; for example, members of the kinds *zinc* and *gold* have different atomic number, boiling point, melting point etc. As the shared properties of *zinc* and *gold* overlap, however, a hierarchy is formed in which the natural kinds *zinc* and *gold* are subkinds of the natural kind *metal*.

6. **Natural kinds should be categorically distinct:** The boundaries between two different natural kinds need to be distinct and not allow for any uncertainty as to how they differ from any other natural kind (Ellis 2001, 99). For example, there are no atoms which are intermediate between *Hydrogen, H* and *Helium, He* by possessing 1.5 protons in the nucleus.

It is a matter of debate whether all the above characteristics should be necessarily associated with natural kinds. According to the ‘metaphysical view’ all the conditions listed above are necessary and (jointly) sufficient for natural kind status. On the other hand, according to the ‘epistemological view’ the only (and, as such, sufficient) necessary condition is that natural kinds should permit successful induction (condition 2).

It seems to be the case that the only condition accepted by both sides as a necessary condition is the aptness for successful induction and explanation. This aligns with the fact that the use of natural kind terms seems to be particularly frequent in scientific disciplines, which operate predominantly via successful induction and explanation. Examples of such induction and explanation are the following:

- Silver has a melting point of 961.72°C. (example of induction)
- If *x* is a sample of silver, then *x* has a boiling point of 2,162°C. (example of induction)
- Silver is used for cutlery because of its antibacterial properties; it irreversibly damages key enzyme systems in the cell membranes of pathogens. (example of explanation)
We are able to make successful inductions on the basis of the fact that samples of the kind *silver* share common properties such as melting point, boiling point and so on. This has been established through the observations made on several samples of silver which were then projected successfully to all members of the category. In addition, we are able to give an explanation of the reason we use silver cutlery by referring to the kind *silver*. Because of the aptness shown for successful induction and explanation such categories are considered epistemically useful; they facilitate understanding and increase our knowledge of several phenomena. Millikan claims that science could not even begin unless we can make numerous generalisations for a single collection of things. She states:

> No science consists of a single generalisation, nor of a heap of generalisations about different kinds of things. A science begins only when, at minimum, a number of generalisations can be made over instances of a single kind, for example over instances of silver, or instances of humans, or instances of massive bodies, or instances of, say, moments in the American economy. (Millikan 1999, 48)

As Millikan claims above, these categories play a fundamental role in scientific practice and referring to them forms an essential part of the methodology of scientific disciplines.

In what follows I discuss how the two different views explain the connection between natural kinds and an aptness for successful induction and explanation. I first look at essentialism, that is, the metaphysical view that natural kinds have an intrinsic essence, a property that is essential for membership into the kind and which explains the laws of nature that reference to the kind reveals (Ellis 2001, 150). I reject the view that natural kinds are metaphysically fundamental because they possess unchanging intrinsic essences which we discover exclusively *a posteriori*. I then turn to recent accounts falling under what I call the epistemological view. According to these accounts, natural kinds just are clusters of properties whose co-instantiation reveals patterns of causal relations (Boyd 1991; Khalidi 2013) or which show stability (Slater 2015). I conclude that Slater’s account provides a plausible account of natural kinds by accepting condition 2 (aptness shown for successful induction and explanation) as the only necessary condition for natural kind status and rejecting all other conditions.
2.3 Metaphysical view: Essentialism and natural kinds

According to essentialism, natural kinds have intrinsic essences; these are essential properties the possession of which is a necessary condition for membership of the kind and explains all other properties possessed by kind members (Bird 2009, 7). Essentialist views received support from Putnam’s insights on how to determine the reference of a speaker’s terms (Putnam 1973). Putnam claims that determining the reference of a kind term can depend on facts external to the speaker herself. According to this view, the meaning and reference of some of the words we use is not solely determined by the ideas we associate with them or by our internal physical or mental states. Instead, the meaning and reference are determined by intrinsic essences possessed by objects which we discover \textit{a posteriori}.

Putnam uses the Twin Earth thought experiment to conclude that we intuitively expect kinds of things to have intrinsic essences, which determine the properties exhibited by objects (Putnam 1973, 226). The Twin Earth thought experiment asks us to imagine a planet, Twin Earth, which is exactly like Earth in all aspects. A twin equivalent object or person exists on Twin Earth for each object or person existing on Earth. The only difference is that a liquid called ‘water’ on Twin Earth is not $H_2O$. All manifest properties of water and ‘twin-water’ are the same; both are colourless, tasteless, odourless, drinkable and so on. However, the molecules of ‘twin water’ are not $H_2O$ molecules; instead they are XYZ. Putnam claims that although speakers on Earth and Twin Earth use the same term ‘water’ this term refers to two different things only one of which is the essence of water and that is $H_2O$ (1973, 224). Even if the speakers have similar psychological states when they are using the term ‘water’ what they refer to depends on the underlying microstructure of the liquid they are talking about; in other words the thing they are referring to does not depend on the meaning they assign to it but on the intrinsic essence of the thing; in the case of water its intrinsic essence is its underlying microstructure (1973, 225). According to Putnam, Twin Earth speakers and Earth speakers do not refer to the same thing when they use the term ‘water’.

Putnam’s thought experiment provides support for the claim that what makes some kinds of things what they are does not depend on our descriptions of them, or on prior uses of the terms associated with them but on intrinsic essential properties possessed by the kinds of things themselves. Putnam’s claim is supported by psychological experiments involving
adults and children which suggest that people take the unity of many kinds to rest on ‘hidden essences’ that are unknown to them but can be discovered through science (Millikan forthcoming, 3). According to this view, we intuitively think that things are the way they are because they possess an intrinsic essence. Possession of this essence is a necessary condition for the thing being what it is. According to Putnam (1973), we establish this essence *a posteriori*, that is, we determine what the essential properties are empirically. In Putnam’s example it turns out upon empirical investigation that what makes ‘this liquid’ *water*, is *being H₂O*. Essentialist accounts claim that the kinds described by Putnam’s semantic externalism are the natural kinds.

According to the metaphysical view of natural kinds, the possession of an intrinsic essence or intrinsic essential properties is what explains the aptness of the category for induction and explanation. These essential properties are shared between all members of the kind and are responsible for other properties exhibited by the kind. Intrinsic essences reveal laws of nature on the basis of which we can create successful induction and explanation relating to the members of natural kinds. For example, according to this view, it is a law of nature that all silver atoms boil at 2162°C – and this is so because having 47 protons in the atomic nucleus is associated with the property of boiling at 2162°C. *Having atomic number 47* is *silver*’s intrinsic essence. Thus, the intrinsic essence of *silver* is what guarantees successful induction and explanation.

According to essentialism, all the conditions mentioned in the previous section are necessary conditions for natural kind status (Bird 2009). As these conditions are considered necessary, kinds that do not satisfy them cannot be considered natural kinds. It is possible that some artefact kinds such as *chair* also satisfy most of the conditions on the list given above. However, according to essentialism, condition 4 (that members of a natural kind should form a kind) places a constraint on kindhood such that only groupings possessing an intrinsic essence qualify as kinds (Ellis 2001). As mentioned above, this essence is discovered *a posteriori*. This is supposed to guarantee that since intrinsic essences are only discovered *a posteriori*, we cannot know them *a priori*. If the intrinsic essence was knowable *a priori*, then one could claim that the classification was imposed on nature by our stipulations rather than by independently existing ‘joints in nature’. Therefore condition 4 may be used to exclude gerrymandered groupings from qualifying as kinds. To give an example, according to the metaphysical view the group *lampsshade* would not qualify as a kind because there is no intrinsic essential property possessed by all members
of the group lampshade e.g. they are made from different materials, they do not share any intrinsic essential property which is associated with the possession of other properties, they have different shapes and uses and so on.

I shall discuss two criticisms of the metaphysical view. First, as LaPorte has argued, discovering essential properties is not exclusively an *a posteriori* process (LaPorte 2004, 110). The classification process is influenced by beliefs, interests and preferences attaching to the theoretical background within which the classification process is taking place. Second, most of the supposedly necessary conditions associated with natural kinds and more specifically the claim that natural kinds possess intrinsic essences, seem to derive from the examination of very few examples of kinds in chemistry or physics. It is not clear on what grounds we presuppose that these kinds are the paradigmatically ‘natural’ ones and that all conditions associated with those kinds are necessary for all natural kinds. Had we chosen to look at different kinds such as biological kinds, or at a conjunction of kinds including physical, chemical and biological kinds, the conditions we would have associated with natural kinds would have been different. I conclude that these criticisms cast doubt on claims associated with the metaphysical view and that we have reasons to think that the epistemological view provides a better account of natural kinds and their relation to successful induction and explanation.

### 2.3.1 Discovering essences exclusively *a posteriori*

Although according to essentialism we discover the essential properties of a kind *a posteriori*, LaPorte (2004, 103-111) claims that this does not accurately describe the process by which we ‘discover’ the essence of an object. He argues that there are cases, such as the case of water, where the natural kind term’s reference is at least in part a matter of stipulation rather than discovery. ‘Water’ can, and in fact often does, refer to mixtures of H₂O and D₂O. D₂O (deuterium oxide ^2H₂O) is a form of water (also called heavy water) that contains a larger than normal amount of the hydrogen isotope Deuterium D, which does not have identical chemical properties with ^1H₂O (protium oxide). Heavy water is rare and occurs in small proportions in samples which are predominantly ^1H₂O. When Deuterium was discovered, in 1931, it was placed in the same location as Hydrogen on the periodic table because it has the same atomic number with Hydrogen. Scientists were then faced with the classification of several substances which included Deuterium such as the substance D₂O. Although D₂O has a different molecular structure to ^1H₂O, both
substances are called ‘water’ even though they have different chemical properties e.g. when a large fraction of water (> 50%) in higher organisms is replaced by heavy water, the result is cell dysfunction and death (Kushner et al. 1999). Whilst in the process of discovering the intrinsic essence of water and classifying substances accordingly, chemists had at least three possible choices relating to the extension of ‘water’: Option 1: water is $^1\text{H}_2\text{O}$ only. The substance of $\text{D}_2\text{O}$ (heavy water) is not water. Option 2: water is the disjunction of $^1\text{H}_2\text{O}$ + $\text{D}_2\text{O}$. If something is only $^1\text{H}_2\text{O}$ or only $\text{D}_2\text{O}$, it is not water. Option 3: water is $^1\text{H}_2\text{O}$ or $\text{D}_2\text{O}$ or $^1\text{H}_2\text{O}$ + $\text{D}_2\text{O}$. The substances in option 3 could be considered as subkinds of the kind water. Chemists chose the third option and included both substances under the classification of water as $\text{H}_2\text{O}$. What this shows for LaPorte, is that science did not discover that water is $\text{H}_2\text{O}$. Chemists could have agreed that the extension of ‘water’ was only $^1\text{H}_2\text{O}$, i.e. the compound made with protium only, on the basis that ‘it bears the relation same microstructural kind to the majority of what we called ‘water’” (LaPorte 2004, 106).

LaPorte claims that in this case the intrinsic essence of water was not revealed exclusively by pure discovery; it also included a proportion of stipulation which was relevant to the interests and classification purposes of chemists. Classifications of substances such as $^1\text{H}_2\text{O}$ or $\text{D}_2\text{O}$ follows the periodic table divisions which are made on the basis of atomic number regardless of whether these substances have different chemical properties e.g. heavy water can be poisonous whereas water is not. As such, when determining whether both substances would be included as extensions of water, some of the chemical properties of those substances were not considered as important as other properties. This importance depends on the interests and preferences of chemists e.g. potability is not considered an important difference in this case. Samples of water will contain molecules of both substances in different proportions. However all qualify as water.

This criticism puts pressure on the essentialist claim that intrinsic essences are a matter of pure discovery and that no stipulation is involved in the process of classification. This claim is used to support essentialist accounts in that it supposedly removes elements of conventionalism, stipulations or indeterminacy from the classification process. As such, kinds with intrinsic essences are metaphysically fundamental in comparison to kinds which lack intrinsic essences. On the other hand, such conventionalism and indeterminacy are associated with gerrymandered or artefact kinds, according to this view. However, it seems to be the case that such stipulation is involved in the classifications
made in chemistry whose classifications are considered to be prime examples of natural kinds. Although determining intrinsic essences via pure discovery is supposed to prove that a kind is a natural kind rather than a gerrymandered kind, it seems to be the case that this process is influenced by considerations of a different type. LaPorte concludes that the case of water shows that ‘the orthodoxy that Water=H₂O was discovered to be true by empirical investigation seems to be wrong’ (LaPorte 2004, 110). The reasoning behind his conclusion is that there was another candidate namely D₂O that could be considered identical to what we call water on the basis of resembling the majority of the samples of liquid that represent the kind water. He states:

\[
\text{H}_2\text{O could not have been discovered to be the essence of what we have all along called ‘water’ because there was another candidate, an overlapping but distinct kind that might as well have been said to be identical to what we called ‘water’, by virtue of being relevantly like the majority of samples of matter representing the kind.} \quad \text{(LaPorte 2004, 108)}
\]

LaPorte does not reject the claim that there could be statements about intrinsic essences that are discovered to be true by empirical investigation alone (LaPorte 2004, 111). However, he does not think that this happens in all cases or that this condition should be used to distinguish natural kinds from gerrymandered kinds. Another important consideration resulting from this discussion is that the existence of conventionalism, or dependency on the classificatory purposes of a theoretical domain, does not entail that the kind cannot be considered a natural kind.\(^{24}\)

### 2.3.2 Excluding kinds without intrinsic essences

According to essentialism, kinds with intrinsic essences are natural kinds; kinds that lack intrinsic essences are not natural kinds (Ellis 2002, 154). For example the chair in my living room and the chair in my kitchen do not share an intrinsic essence as they do not possess an intrinsic property which is a necessary condition for being a chair. Although they share superficial properties such as their appearance and perhaps use, according to essentialism there is no underlying essential property shared by both chairs; for example,

---

\(^{24}\) This claim is important for the view that social kinds can be considered natural kinds as I argue in chapter 3. Theory dependency or conventionalism are used as objections to the claim that social kinds can be natural kinds. The discussion in this chapter shows that these are not necessarily incompatible with natural kind status, unless one holds that chemical kinds are not natural kinds.
one might be made of wood and the other of metal or plastic. As such, the kind chair is not a natural kind. According to this view, having an identical underlying essence explains the existence of all other common properties exhibited by the kind members and, as such, explains why these kinds show an aptness for induction and explanation.

The conditions mentioned in the previous section as necessary for natural kind status appear to be associated with specific types of categories in chemistry and physics such as chemical elements or microphysical particles. However, it does not seem likely that they can be applied to other scientific disciplines dealing with more complex entities such as biological or social sciences. For example, if the common essential properties are viewed as intrinsic it would be very difficult to reconcile condition 1 (that natural kinds should share natural properties) with the ever-changing properties characteristic of biological species. As LaPorte or Millikan argue, it is possible that biological kinds have historical essential properties which are not intrinsic (see LaPorte 2004, or Millikan 1999 for more detail). This would mean that biological species have historical essences. I do not discuss this view as I focus on essentialist views which view essence as intrinsic such as Ellis’ ‘new essentialism’ (2001; 2002).

Ellis claims that because biological kinds do not possess intrinsic essences, they are not natural kinds (Ellis 2002, 154). He states that:

According to the new essentialism, the true natural kinds of substances exist only at a much deeper level than that of living species. They include the basic kinds of physical and chemical substances, such as the various species of atoms, molecules and subatomic particles, but not the biological kinds. (Ellis 2002, 12)

According to Ellis, if a kind lacks an intrinsic essence then it is not a natural kind. According to such a view, there seem to be very few natural kinds out there in the world. These are restricted to the very few kinds whose members share an intrinsic essence, such as the physical or chemical kinds. This claim takes away most of the special epistemic status traditionally associated with scientific categories in biology, psychology, sociology and so on as it implies that the categories we use in all these scientific disciplines are not natural kinds.
However what is an issue for essentialism is that there are many biological kinds such as *tiger, dolphin, oak*, which show a high degree of aptness for induction and explanation. In these cases we lack an explanation for the fact that these categories show an aptness for successful induction and explanation. This is because, according to essentialism, the possession of an intrinsic essence is what explains the successful induction and explanation associated with the category. However, these categories do not possess intrinsic essences. As they do show an aptness for induction and explanation, one can claim that the possession of an intrinsic essence is not necessary for showing an aptness for induction and explanation.

In addition, it is not clear why possessing an intrinsic essence is a necessary condition for natural kind status. To determine what conditions are necessary for natural kind status we need to investigate different types of kinds to see what, if anything, is shared between them. It seems arbitrary to use kinds exclusively from chemistry and physics in this process as this entails that we are presupposing that these are the only candidates for natural kind status. However, it is not clear that we have to make this presupposition. It is one thing to claim that the kinds *silver* and *gold* are in some sense more natural than the kinds *canine* and *oak tree*, but it is a completely different thing to claim that the kinds *canine* and *oak tree* are not natural at all. That would require us to first make some sort of metaphysical commitment concerning what counts as ‘natural’. By contrast, if one does not subscribe *a priori* to any metaphysical commitments as to what counts as natural, then determining the conditions for natural kind status will be a result of the examination (*a posteriori*) of what is shared between different types of kinds in nature. As nature includes biological or social kinds, such candidate kinds will also be included in this examination.

If we compare kinds from all different disciplines and domains it becomes apparent that what is different between kinds such as *canine and dog walker* is a difference in the aptness shown for successful induction and explanation. All other conditions listed in the previous section, and suggested as necessary by essentialists, seem to be the result of focusing on specific types of candidate kinds whilst excluding others. In the next section I discuss alternative accounts falling under the epistemological view, which reject the main thesis of essentialism that what is shared between members of all natural kinds is an intrinsic essence in the sense of intrinsic or microstructural properties.
2.4 Epistemological view: Property Clusters

According to the epistemological view, what separates natural kinds from non-natural kinds is the aptness shown for successful induction and explanation. For example consider the kind *dog walker*. It seems unlikely that referring to this kind will lead to successful induction and explanation relating to dog walkers. Anyone can be a member of such a kind (apart from some exceptions such as people with dog phobias and so on) and it is not likely that observations made on some members of the kind can reveal properties shared by other members of the kind such as age, height, weight, fitness, ethnicity, behaviour and so on. One could argue that we could make some induction and explanation about members of the kind *dog walker* such as that ‘most dog walkers are not afraid of dogs’; however, such induction is not likely to be as numerous as the one relating to the kind *silver* or reveal any further properties of the kind members. The kinds *silver* and *dog walker* appear to be different at least in this respect, then: the former shows an aptness for successful induction and explanation whilst the latter does not or shows a very limited capacity for induction and explanation. Therefore it seems likely that the aptness shown for induction and explanation comes in degrees; some kinds show greater aptness than others, whilst it is possible that some gerrymandered kinds such as a collection of random objects located in my living room at a certain point in time shows very little or none at all. However it seems to be the case that this gradation is missed completely under the essentialist view. According to essentialism, kinds either do or do not possess an intrinsic underlying essence. There are no intermediary cases or cases where one natural kind will be more natural than another natural kind.

In what follows I separate the epistemological view into two broad categories: 1) causal property clusters (CPC) and 2) stable property cluster (SPC) accounts. CPC accounts consider causation and causal relations as the ultimate basis of classification. By contrast, according to the SPC account, although some natural kind classifications may reveal causal relations there are others that do not. According to the SPC account what drives the classification is the stability exhibited by clusters of properties – but SPC does not require that such stability is explained in causal terms. The accounts falling under the epistemological view take naturalness to depend on the aptness for successful
generalisation and induction; there are many categories in the natural and special sciences that are candidates for natural kind status on the basis of this criterion and, as a result, natural kind status is not restricted to a few chemical and microphysical categories. I argue that these views provide better explanations for the aptness shown by some categories for successful induction and explanation regardless of the possession of intrinsic essential properties.

All of the accounts to be discussed reject the claim that natural kinds possess intrinsic essences in anything like the essentialist sense of intrinsic essence. According to CPCs natural-kind classification reveals causal patterns or relations. It is these causal relations that explain the aptness shown for induction and explanation. Although such accounts are more plausible than the metaphysical view, I argue that they replace a necessity for ‘essences’ with a necessity for causal relations. However, as we have kinds who show an aptness for induction and explanation without revealing any causal patterns, such accounts should not be used as the basis for a plausible account of natural kinds. Instead Slater’s account replaces a necessity for causal relations with a necessity for stability portrayed by the clustered properties. It is the stability which can explain and determine the degree of successful induction and explanation.

2.5 Causal Property Clusters (CPCs) and Homeostatic Property Clusters (HPC)

An influential account suggested by Boyd argues that we should assume a kind to be successful (natural) when it reveals to us the causal architecture of nature (1999a). According to this account, it is because natural kinds reveal causal structures, in the sense of revealing patterns of causal relations between properties, that they are able to generate successful induction and explanation about kind members.

Boyd argues that the members of some natural kinds share Homeostatic Property Clusters (HPCs), clusters of properties co-instantiated and maintained because of the existence of certain causal mechanisms. These mechanisms are responsible for the maintenance of

---

25 Several philosophers suggest that naturalness should be determined on the basis of whether the kind allows for successful generalisations and predictions within the background of a particular theoretical domain (Griffiths 1997; Boyd 1991; Khalidi 2013; Mallon 2016).

26 Boyd does not claim that all natural kinds are HPCs but that some natural kinds are. His stronger claim is that all natural kinds reveal patterns of causal relations between properties. In the second section of this chapter I give reasons why even this stronger claim can be applied to some natural kinds but not all; in other words I argue that some natural kinds do not reveal patterns of causal
the properties within the cluster without any of the clustered properties being necessary or sufficient for membership of the kind. Boyd uses the example of a biological species as an example of a HPC. He claims that members of a biological species share homeostatic mechanisms (intrinsic or extrinsic) that are responsible for the collection and perseverance of certain sets of properties:

Birds of prey are natural kinds defined by the properties that species’ member organisms typically exhibit in combination with the mechanisms that underwrite this clustering such as common descent, reproductive cohesion or environmental factors. (Boyd 1999, 72)

Living organisms share sets of properties that cluster together over time because of several intrinsic and extrinsic factors. Any group of organisms sharing these clusters could be considered a natural kind. For example, on this view, the HPC dolphin is a natural kind. We are able to observe similar properties that are common to most dolphins such as that they have hairless bodies, they are fishlike in form, they breathe air through a single, dorsal blowhole, they have tailfins, fore fins and a dorsal fin, they have a layer of blubber to protect them from cold, the female gestation period lasts 12 months, they use an echolocation mechanism to communicate with each other and to determine size, location and nature of surrounding objects, they are capable of night vision, they have a thermoregulation mechanism and so on (Jefferson et al. 2008).

It is not clear according to this account whether the homeostatic mechanisms are necessary or sufficient for membership of the kind or whether one homeostatic mechanism is singly responsible for the clustering of the properties. In the example of the kind dolphin we could think of several homeostatic mechanisms which are responsible for and maintain the clustering such as evolutionary history (extrinsic), thermoregulation (intrinsic), osmoregulation (intrinsic) and so on. As Boyd claims that there are no necessary and sufficient conditions for membership of an HPC, it is fair to assume that,

relations and as such, this should not be used as the criterion on the basis of which we determine ‘natural kindness’.

27 Boyd claims that the clustered properties are maintained by homeostatic mechanisms or homeostatic properties; however it is not clear what ‘homeostatic properties’ are or whether they are included in the property cluster or they are outside the property cluster. I will focus on homeostatic mechanisms for the remainder of this chapter.

28 Osmoregulation controls the water concentration in blood and body fluids. Failure to regulate water concentration leads to failure of vital internal organs such as the kidneys and may cause dehydration and death.
according to this view, no single homeostatic mechanism is necessary and sufficient for membership into the kind. In the passage quoted above, Boyd claims that the kind is defined by the sum of the clustered properties plus the mechanisms (intrinsic and extrinsic) that underwrite the clustering. As such, it appears that he considers all the relevant mechanisms to be jointly responsible for the clustering and maintenance of the properties.29

To give an example of how a homeostatic mechanism works, consider thermoregulation. Dolphins require a constant body temperature of 36-38°C, similarly to other mammals. As dolphins’ metabolism is higher than that of other mammals of similar size, it generates a great deal of heat, which exceeds the designated temperature of 36-38°C. On the other hand as they dive to great depths the temperature of the environment falls dramatically, which causes their body temperature to fall below the designated temperature. Dolphins regulate their body temperature in any of the following ways (Jefferson et al. 2008):

- They have a blubber which insulates against heat loss
- They release heat through the dorsal fin and flukes
- During dives they reduce blood circulation to concentrate at the core organs/tissues rather than the outer parts of the body and so preserve heat
- They move to cooler/warmer areas
- They transfer heat from one blood vessel to the other without releasing it to the environment.

According to Boyd, biological kinds that define membership on the basis of homeostatic property clusters have a flexibility in the sense that they allow for change and variation within members of the same kind. At the same time Boyd claims that biological kinds are paradigmatic natural kinds because we are able to make numerous significant inductions and explanations about members of species. To give a few examples, by referring to biological kinds we can make the following inductions and explanations:

- All dolphins have a thick layer of blubber.

29 Millikan disagrees on this point and argues that common descent should be considered as the most important mechanism for the clustering of the properties. See Millikan 1999 and Boyd’s reply 1999b.
- Dolphins have a dorsal fin to prevent involuntary spinning in the water.
- Dolphins can travel at speeds up to 34.5 m/h.
- The dolphin reached the fish before the seal did because dolphins swim faster than seals.

Boyd’s view allows that none of the clustered properties on its own is necessary or sufficient to determine whether an organism should be included in the kind *dolphin*. Some or any of the properties included in this cluster may be missing in individual cases and other organisms can share some of the properties in the cluster without being dolphins. For example sharks have tailfins, fore fins and a dorsal fin, whereas some dolphins do not have fore fins. However, an absence of certain properties in such individual cases would be the exception and these absences would not persist with enough frequency in all members of the kind so that the cluster of properties is permanently altered. Therefore, although the clustered properties within HPCs show a stability, individual properties within the cluster may change over time.

To give an example, consider the property of having a prehensile tail in primates. For a long period of time, this was a very important property for primates because they were forced to live in trees for survival reasons. As soon as environmental factors changed and it became safe to inhabit open land areas, this property did not persist. It seems likely that this property, which was included in the HPC of the kind *primate*, disappeared when external factors changed. In this example, the external factors acted as the causal mechanism which maintained the property for a period of time. Once the external factors changed, the property disappeared from the cluster. This means that this account rejects condition 1 relating to sharing intrinsic properties as necessary for membership as any of the clustered properties may be missing or may change. This is in direct opposition to the essentialist claim that all members of natural kinds share an intrinsic essence in the sense of intrinsic essential properties whose possession is necessary for membership.

According to Boyd, all natural kinds are apt to generate successful inductions and explanations. This leaves the door open for innumerable kinds to qualify as ‘natural’, as long as they satisfy this condition (condition 2). According to Boyd’s view:

> The naturalness of a natural kind is a matter of the contribution that reference to it makes to the satisfaction of the accommodation demands of a
disciplinary matrix, in the context of a system of compositional linguistic resources for the representation of phenomena. (Boyd 1999a, 158)

To put this in simpler terms, to determine whether a kind is natural we have to look at the contribution that reference to the kind makes to a particular discipline. If we are able to make successful predictions and explanations by referring to this kind, it is a natural kind according to Boyd. Trying to explain this claim, he states that ‘the naturalness … will ordinarily be a matter of the role that reference to it (the kind) plays in some particular family of inductive or explanatory practices’ (Boyd 1999a, 158). Therefore the naturalness of the kind is determined on the basis of the role being performed by reference to this kind within a specific theoretical background. Boyd refers to this process as the ‘Accommodation condition’, that is, the process of matching the inferential practices within a specific discipline with the relevant causal structures in nature (Boyd 2010, 215). A discipline uses categories to try and make successful inductions and explanations about the phenomena it investigates. When these inductions and explanations are successful, because they reveal the causal relations that exist in nature, these kinds can be considered natural.

According to Boyd only certain property clusters qualify as natural kinds. The kind dog walker is not likely to be considered a natural kind because any properties shared between the members will allow very few, if any, successful inductions or explanations. For example it is not likely that we will be able to predict the height, weight, age or fitness levels of the category after observing a sample of dog walkers. There are no rules which need to be satisfied so that someone becomes a dog walker by way of tests, qualifications and so on. The hypothetical cluster of properties shared by members of this kind is not likely to be the basis for important induction and explanation in the sense suggested by Boyd. As Boyd is particularly interested in patterns of causal relations between the clustered properties it is not likely to be the case that such patterns will be discovered for the members of the kind dog walker because a) it is not likely that there will be enough clustering of properties apart from the property of being able to take a dog for a walk and b) there is no common underlying causal mechanism responsible for that ability e.g. one can be fit or unfit, walk or ride a bike, one can be a teenager or a child and so on.

Causal kind accounts such as the HPC account take causal relations to be responsible for the properties that cluster together. Boyd claims that the properties which cluster together
reveal the causal structure of nature. The assumption here is that nature has a causal structure to be revealed by our epistemic projects. According to such views, it is the correspondence between the generalisations made by a particular domain with the causal structure of the world that is supposed to provide the connection with realism (Boyd 1991, 2010; Dupré 1993; Griffiths 1997; Craver 2009; Khalidi 2013; Millikan, forthcoming). Boyd argues that:

The causal importance of the homeostatic property cluster $F$, together with the relevant underlying homeostatic mechanisms, is such that the kind or property denoted by term $t$ is a natural kind. (Boyd 1999a, 144)

Causality is therefore a very central element by being responsible for the presence of specific properties and mechanisms within the HPC. According to this view, a cluster of properties, such as A, B and C are frequently co-instantiated. It could be the case that property A caused properties B and C, or property A is responsible for the presence of property B, which in turn caused property C, or that A, B and C are all individually caused by a separate property or mechanism. Either way, a causally important relation is responsible for the co-occurrence of these properties (Boyd 1999a, 143). To use the example of dolphins mentioned in the previous section, the homeostatic mechanisms of common descent, reproductive cohesion and environmental factors are causally important for the maintenance of the HPC dolphin.

According to the HPC view, cases may arise where determining membership will be difficult. Boyd does not consider this to be a problem because he takes the causal architecture of nature to be messy. He uses biological species as a paradigm of a HPC precisely because there can be indeterminate species or difficult cases, for example before the creation of a new species or after the union of two different species. Consider the example of mules. The mule is a product of the reproduction between a donkey and a horse. Mules do not share enough properties from the HPCs associated with the kind donkey or the kind horse to count as a member of either of those natural kinds. As such, the kind mule is considered a biological kind that is separate to the kinds it descended from. Biologists make successful inductions and explanations for the kind such as the following:

- Mules have 63 chromosomes.
• Most mules are infertile because the number and different structure of chromosomes usually prevents their proper pairing up so that successful embryos are not created.

With respect to the genotype, mules have 63 chromosomes, which is a mixture between the 64 chromosomes of the horse and the 62 chromosomes of the donkey. The mule’s HPC has similarities with both the HPCs of *horse* and *donkey*, however the mule does not share enough similarities with either to be considered a donkey or a horse. Although the mule shares many properties with the kinds *horse* and *donkey*, we cannot determine whether it qualifies as a member of either kind because there are no necessary and sufficient properties for membership into the kinds *horse* or kind *donkey*. As it also has properties which neither of its ancestors possess, biologists categorise it as a hybrid. This view posits a messy ‘architecture’ of nature as it allows for cases where it is difficult to determine boundaries between kinds. Nevertheless, hybrids exist in nature regardless of the epistemological fact of whether we discover which kind they belong to. This account appears to be more interested in highlighting the causal relations between properties and mechanisms and to a lesser degree in setting the boundaries of the cluster. The latter is a matter to be dealt with, where possible, *a posteriori* by our best possible science. As such this accounts rejects the conditions 5 and 6 mentioned in section 1 relating to hierarchy and clear and distinct boundaries.

The HPC account accepts condition 2 (that natural kinds should permit successful induction) as a necessary condition for natural kinds. It could also be argued that the HPC account could accept condition 3 (revealing laws of nature), if we understand laws of nature as regularly occurring patterns of causal relations. According to the HPC account all other conditions (apart from conditions 2 and 3) associated with natural kinds are not necessary conditions for natural kind status. In the next section I look at a particular criticism of the HPC account raised by Craver (2009) and associated implications attaching to the application of this account to the psychological or social sciences.
2.5.1 Property clusters and mechanisms

According to Boyd, biological kinds are candidates for natural kind status in the same way as microphysical and chemical kinds. Certain other kinds such as psychiatric, psychological or social kinds also show an aptness for successful induction and explanation and, as such, could be considered as candidates for natural kind status under the epistemological view. However, Craver argues that in some sciences such as psychology, HPC’s reliance on mechanisms invites conventionalism into the process of classifying, moving this account further away from realism (Craver 2009). According to the HPC account, the existence of homeostatic mechanisms is supposed to explain the co-instantiation and maintenance of the properties. However, Craver argues that defining and individuating mechanisms depends on our interests and classificatory purposes. As such, it is not clear that searching for homeostatic mechanisms adds an objective constraint to the classification:

Conventional elements are involved partly but ineliminably in deciding which mechanisms define kinds, for deciding when two mechanisms are mechanisms of the same type, and for deciding where one particular mechanism ends and another begins. This intrusion of conventional perspective into the idea of a mechanism raises doubts as to whether the HPC view is sufficiently free of conventional elements to serve as an objective arbiter in scientific disputes about what the kinds of the special sciences should be. (Craver 2009, 575)

Craver’s claim is directed against a strong interpretation of HPC theory according to which distinct homeostatic mechanisms correspond to distinct natural kinds (Craver 2009, 580). According to such a view for example, if memory or emotion are natural kinds there should be a single causal mechanism underlying each kind. However, as Craver claims, to differentiate and define mechanisms we need to consider the purposes of the classification, our interests and the phenomena under investigation. Craver claims that in the special sciences such as in psychology:

The boundaries of a mechanism depend on our choice of the explanandum phenomenon (that is our choice of the property cluster) and on the way we choose to describe that phenomenon or property cluster. If so, then the mechanistic structure of the world depends in part upon our explanatory interests and our descriptive choices at the level of property clusters…..there are
An example used by Craver relates to the psychological construct of memory. The hippocampus is a brain region thought to house a mechanism involved in encoding declarative and spatial memories. Humans and other mammals have one hippocampus in each side of the brain. According to Craver, when examining specimens of the hippocampus region scientists noted that:

If one focuses on the particular cells and structures in these specimens, their particular locations and shapes, their particular activities, their exact numbers, and so on, then no two hippocampi are identical. Even those in the same head or in the same location moment to moment are different in many respects. If the HPC account were to suggest that we should split kinds when the mechanisms differ in any of the myriad ways that any two mechanisms might differ – in the precise features of their component entities, activities, and organization – then there would be as many kinds of declarative memory encoding mechanism as there are individual hippocampi. (Craver 2009, 585-586)

A person who is interested in creating a classification of memory phenomena may look to the underlying causal mechanisms to individuate different categories. As Craver claims, to individuate one mechanism from another one can choose to use different criteria such as activities performed, organisational structure, physical composition and so on. One can choose to group all these mechanisms together under one large group on the basis of the fact that they are performing similar activities ignoring the differences between the physical aspects of the mechanisms or the differences in organisation (Craver 2009, 587).

On the other hand if we define the mechanism in maximal detail by focusing on the physical aspects, then each individual mechanism would be a kind in itself, which ignores similarities in activities performed by these different mechanisms. Ultimately our decision depends on what our interests are and which level of abstraction we choose to investigate. In the example of the hippocampus used above by Craver, if we are interested in differentiating memory mechanisms from non-memory mechanisms we can lump different mechanisms together on the basis of type of activities performed. If we are...
interested in differentiating different types of memory we will split the different mechanisms to reflect the difference in the type of memory activities performed or the type of organisation or physical structure and so on. In any case how we group these mechanisms depends on a prior decision of which phenomena we choose to investigate.

According to Boyd’s account, the presence of a homeostatic mechanism is supposed to explain the clustering of the properties. However what counts as the same mechanism or the same kind of mechanism depends on our interests. For example if we think that declarative memory is a useful category to have, we consider the mechanism underlying declarative memory to be the same in all people. However, if we just focus on the mechanism itself without relating it to any category, there is no principled way to decide whether the mechanism is the same in each person at different times or the same in different people. This implies that the existence of the mechanism is not providing any additional objective grounding which explains why the category of declarative memory is a useful category to have. To give another example we can ask whether echolocation is a natural kind. The mechanisms responsible for echolocation in dolphins and bats have both similarities and differences, therefore it is not clear that we can provide an answer just by looking at the mechanisms involved. If Craver is right then the only way to settle this question is to decide first whether echolocation is a useful category to have.

Craver concludes that the appeal to homeostatic mechanisms invites conventionalism into the process of classification, rather than providing an objective criterion which we could use to distinguish kinds. He states:

The defender of the HPC view appeals to mechanisms in order to explain why a given property cluster is useful for the purposes of prediction, explanation, and control. But this promiscuous strategy reverses the direction of fit. It accepts mechanism schemata of different degrees of abstraction because they are useful for the purposes of prediction, explanation, and control. At this point what work is the appeal to mechanisms doing for the scientific HPC? A simple causal view of natural kinds could fill this role well. (Craver 2009, 580)

Craver claims that there are two possible ways to provide a reply to this issue. First he claims that defenders of the HPC account could find ‘an objective basis for taxonomizing the mechanistic structure of the world’ (Craver 2009, 591). Alternatively, defenders of the
HPC account could develop a more pluralistic view, in which ‘there are many disputes about the taxonomies of the special sciences that will not be resolved’. He explains that in this case a pluralistic HPC view:

Will not tell us in any definitive way how many kinds of memory there are, whether the emotions... are natural kinds, or how to refine the DSM (Diagnostic and Statistical Manual of Mental Disorders). Instead, there will be multiple incompatible answers to these questions depending on which mechanisms one attends to, on how one describes the phenomenon, and on where one draws the boundaries of the mechanism. Perhaps this is as it should be. On a more pluralistic conception of special science kinds, such taxonomic problems are not so much solved, as dissolved. ...If the above considerations are correct, our interests and objectives contribute partly, but ineliminably, to the kinds of mechanisms we find in our world. (Craver 2009, 591-592)

To sum up Craver's view, natural kinds are:

The kinds appearing in generalisations that correctly describe the causal structure of the world regardless of whether a mechanism explains the clustering of properties definitive of the kind. (Craver 2009, 579)

Craver accepts that appealing to mechanisms invites conventionalism into the process of classification, at least in the special sciences. However, he seems to be suggesting that such conventionalism, although inevitable, does not threaten the ‘realist objectives that motivate the belief in the natural kinds in the first place’ (Craver 2009, 591). Natural kinds according to this view, describe the causal structure of the world. This causal structure can be described in different ways which may give incompatible answers to classificatory projects in the special sciences. Craver concludes that this is not problematic if one endorses a pluralistic view of natural kinds and if one rejects the assumption that there must be ‘a uniquely correct and orderly taxonomy of the mechanistic structure of the world’ (Craver 2009, 592). Therefore he accepts that the HPC account invites conventionalism but he does not think that this conventionalism is a threat which we should provide a reply to. Whether or not a property cluster is considered a natural kind depends on whether the generalisations we create on the basis of this cluster correctly describe the causal structure of the world. There is still a constraint in place to the effect
that there is a causal structure which we discover regardless of whether we choose to describe it in different ways e.g. by focusing on different types of memory.

Craver’s account, in a similar way to Boyd, focuses on what he calls ‘the causal structure’ of the world and causal relations which underlie the clustering of the properties (Craver 2009, 592). Khalidi (2013) in a similar way argues that we can give an explanation for the clustering of the properties by appealing to causal relations (2013) rather than homeostatic mechanisms. In the next section I give more detail on how the above researchers view these causal relations and discuss whether these should be used as the criterion on the basis of which we determine natural kinds.

2.5.2 Property clusters and causal relations

Khalidi (2013) argues that the causal relations explaining the clustering of the properties do not need to be thought of as comprising some sort of ‘single causal mechanism’ that is homeostatic. He states:

> When it comes to natural kinds, causal relations among properties (or more properly, property instances) constitute the ontological ground for the projectibility of the corresponding predicates. But there need be no single causal mechanism that leads these properties to be co-instantiated nor need there by any kind of feedback process that ensures that these properties do not depart from an equilibrium state of co-instantiation. So Boyd’s account has to be loosened in such a way as to retain the emphasis on causality without the mechanism or the homeostasis. (Khalidi 2015, 1386)

Khalidi claims that in most instances what explains the co-instantiation of clustered properties is causal structure. This is how Khalidi describes a natural kind:

> Natural kinds are identified with one or more causal properties, which when instantiated or co-instantiated, cause the instantiation of other properties in structured causal networks. The causal links between these properties may not be strict and this gives rise to natural kinds with fuzzy boundaries. (Khalidi 2015, 1380)
Khalidi argues that a natural kind is a set of properties whose co-instantiation causes the instantiation of other properties. According to this view, mere correlation of properties is not enough for a kind to be considered natural; further investigation is needed to determine what grounds this clustering because ultimately we are interested in finding out the causal structure of the world that is the causal links between the clustered properties (Khalidi 2015, 1383). According to Khalidi some of the clustered properties are ‘core’ properties or ‘primary’ properties, which are responsible for the existence of other properties in the cluster, which he calls ‘secondary’ properties. Khalidi gives the following example relating to the natural kind gold:

The primary causal properties of gold include atomic number 79 as well as a disjunction of mass numbers, which give rise in turn to a cluster of other causal properties (such as ionization energies, atomic radius, and so on.). Further when atoms with this cluster of properties are aggregated under certain conditions they give rise to another cluster of properties. The cluster of primary properties corresponds to what I have been calling the ‘core’ properties of this natural kind, while the others are derivative or causally ‘secondary’ (Khalidi 2015, 1384)

Khalidi’s view thus allows for some natural kinds with necessary and sufficient conditions; however he claims that there may be natural kinds where the ‘core properties’ are not singly necessary or jointly sufficient for membership (Khalidi 2015, 1385).

It seems that Khalidi is intending to track the same sort of causal relations as Boyd is. However, he claims that we need not define these causal relations as some sort of mechanism. The reasoning for including the mechanism in Boyd’s account is to increase the objectivity of the account, by providing a common cause responsible for the clustering of the properties. Craver convincingly argues that the appeal to mechanism invites conventionalism because of the difficulties in determining mechanisms. At the same time Craver argues that such conventionalism is not a threat to the realist credentials of the HPC account. On the other hand Khalidi claims that the appeal to mechanisms is not a necessary feature of all natural kinds. There could be natural kinds which do not share any underlying causal mechanism. In Khalidi’s account what is important for explaining the clustering of the properties is the causal relations between the properties in the cluster. As such, although there are differences between these three accounts, all of them focus
on the existence of causal relations that provide an explanation for the clustered properties.

In all the accounts discussed above, the appeal to causality appears to be playing two different roles. The first role, type \(A\), relates to the cause of the co-instantiation of the clustered properties. So type \(A\), relates to tracking the reason why the properties cluster to begin with. Boyd’s account argues that the first role is explained by a homeostatic (causal) mechanism/property. To give an example of type \(A\) causal role, in the case of species, common descent explains the clustering of specific genetic properties. The second type, type \(B\), relates to the role played by the causal relations within the cluster that is the relations between the ‘core properties and the ‘secondary’ properties in Khalidi’s account. For example the property of having atomic number 79 causes the property of having a melting point of 1064°C. \(^{30}\)

Depending on our classification interests we can focus on a different type of causal role (see figure 2 below). So for example in the case of biological species, we can look for the causal relations responsible for the co-instantiation of property cluster \(Y, Z, T\) such as common descent \(X\). Common descent provides the causal reason why properties \(Y, Z\), and \(T\) cluster together. Alternatively, we can look for the different homeostatic mechanisms/properties within the property cluster such as property \(Z\), such as thermoregulation, which is causally related to other properties within the cluster such as property \(T\), such as having a thick layer of blubber. On the basis of the causal relations detailed on figure 2 we can infer statements \(I, J\) such as the statement that ‘all clusters \(YZT\) have a constant body temperature of 36-38°C’.

\(^{30}\) Khalidi argues that the property of having atomic number 79 ‘causes’ the existence of certain other properties in the cluster. It is not clear in what way this relation is one of causation rather than co-existence. However, for the purposes of discussing his account I accept his claim that it could be considered as such.
Khalidi claims that there are many reasons which might explain the co-instantiation of the core properties within the cluster (Khalidi 2015, 1385). These include causal and non-causal factors such as convention, or co-existence lacking any causal explanation. According to this view in some instances, the co-instantiation is a simple co-existence with no underlying causal story or explanation. He states:

At some point, there may be no further causal story to be told concerning the co-instantiation of certain properties, just a constant conjunction of two or more properties that are always (or nearly always) associated in nature. (Khalidi 2015, 1390)

To give an example: the chemical element Tantalum $Ta$ is almost always found together with the chemical element Niobium $Nb$, forming the minerals tantalite, combulite and coltan. There is no causal story that explains why these two elements are always found together in the sense of revealing any causal relations between Niobium and Tantalum.\(^{31}\)

\(^{31}\) Although a historical or geological account of the formation of the Earth may provide additional information it is not likely that such information would be of interest to chemical classifications of
A separate factor which is responsible for the clustering according to Khalidi is convention or decree. He states the following:

Consider social categories like senior citizen, permanent resident, or felon. Roughly speaking individuals who are classified in one of these categories are also classified because they have properties that are conventionally rather than causally associated. In some jurisdictions a person is a senior citizen if and only if one is a citizen and has attained the age of 65, and this is a matter of legislation...When properties are associated in this manner as a matter of legislation, decree, or convention, they cluster together not as a result of causality. (Khalidi 2015, 1393)

Khalidi claims that in these cases we will not discover a causal story which explains the co-instantiation (such as the co-instantiation of citizenship and being over the age of 65 amongst senior citizens). In such cases the properties cluster together not as a result of causality, because such properties are not causally linked, and there are no further properties that are caused by them (Khalidi 2015, 1393) Khalidi claims that such kinds are unlikely candidates for natural kind status. However, he claims that in some social categories causal and non-causal reasons for clustering are entangled, such as the categories of consumer, refugee or psychopath; here, the clustering of the relevant properties is a result of both legislation or convention and causal reasons. According to Khalidi qualifying as a member of the category of refugee has more to do with a person’s participation in certain causal processes or their causal history than with legislation and convention. He states:

For instance, refugees are persons escaping hardship or persecution, they flee or migrate from one society to another, they typically face an adjustment period in the host society, they frequently face discrimination, and so on. These kinds of regularities are ones studied and discussed by social scientists, and they have a causal basis......When it comes to social kinds, the causal and the conventional can be entangled in certain ways. For instance ....permanent resident, is legally or conventionally defined in many
jurisdictions, yet it can also have a causal dimension and come to play a causal role in social processes. (Khalidi 2015, 1394)

Boyd’s account focuses on both type $A$ and type $B$ relations as he allows that there can be homeostatic mechanisms both within and outside the property cluster. Depending on the phenomena under investigation one can focus on any type of causal relations associated with the property cluster. Khalidi claims that there may be cases where even if we do not have a causal story for the clustering of the properties, because these were created by convention or by chance, we can still make successful inductions and explanations on the basis of causal relations between the clustered properties. He claims that several biological or social kinds are good candidates for natural kinds such as cancer, virus, or a specific type of ADHD if we look at what originally caused them or if we look at what is causally determined by having one of those features e.g. how one is being treated if one is a refugee.

According to the accounts mentioned above, natural kinds mostly reveal causal relations associated with clustered properties. Boyd’s and Khalidi’s accounts explain the aptness shown for induction and explanation via the existence of causal relations. Natural kindness in these accounts seems to be tracking causal relations between properties. To sum up: the accounts discussed thus far accept conditions 2 (aptness for successful induction and explanation) and 3 (revealing laws of nature viewed as causal relations). As such these accounts reject conditions 1 (intrinsic essences), 4 (kind formation), 5 (hierarchical structure) and 6 (distinct boundaries).

It is not clear whether the causal relations discussed above are necessary or sufficient for induction and explanation. However, if any property clusters show an aptness for induction and explanation without being associated with any causal relations this would imply that causal relations are not necessarily associated with such an aptness. In the next section I discuss a view according to which not all property clusters reveal causal relations. As such we should not accept that one of the conditions necessarily associated with natural kinds is that they reveal causal relations and as such, we should also reject condition 2.
2.6 Stable Property Clusters; a more plausible account

According to Slater’s view, ‘naturalness’ need not exclusively track the causal factor responsible for the clustering of the properties. Instead, it tracks a stable co-existence of properties regardless of what caused this co-existence. Slater claims that his account differs from Boyd’s account in that it requires only that:

The properties be sufficiently stably co-instantiated to accommodate the inferential and explanatory uses to which particular sciences put such categories.

(Slater 2015, 396)

Slater claims that natural kindness should not exclusively track causal relations and that we should further relax the causal property cluster account such that the status of ‘natural kindness’ is multiply realised (Slater 2015). According to this view, causal relations can be found in some property clusters but not all. For example, according to Slater, it would not be right to claim that kinds of elementary particles reveal causal relations between clustered properties. He states that:

The elementary particles, for example, appear to be individuated by perfectly maintained suites of properties, none of which are derivative from any others. Perhaps it is simply a fundamental law than an up quark has a spin of \(\frac{1}{2}\), charge of \(\frac{2}{3}\), baryon number of \(\frac{1}{3}\), mass of 360 MeV/c\(^2\), and so on. (Slater 2015, 381)

In the quote above Slater, in a similar way to Khalidi, claims that it is not always the case that the properties within the cluster are related by causality in the sense that one derives from another. He claims that we do not seem to have a causal story that explains why these properties cluster together. However such clusters show a stability in that the properties are reliably co-instantiated.

Slater allows for the view that some property clusters may track causal relations whilst others may not, such as elementary particles. Instead he claims that what justifies the success of our inductions and explanations is the stability of the properties in the cluster regardless of whether these are connected by any causal relations. According to this view the stability shown by the cluster properties, viewed as a co-instantiation that is not likely to change, is multiply realised: some clusters show this stability because they have
homeostatic mechanisms while other clusters have properties for which no causal story
can be told (Slater 2015, 398 n.29). This is how Slater describes the concept of stability:

Call this conception of stability ‘cliquish stability’…. The idea is to capture the
fact that some properties are clustered in such a way that possession of
some of them reliably (if imperfectly) indicates the possession of the whole
cluster (if not each property in the cluster) at that time. It need not imply
that a particular that possesses any of the properties will continue to possess
them. (Slater 2015, 397)

Slater seems to suggest that some properties reliably cluster together without this
necessitating that they will cluster together in all instances at all times. Giving more detail
on the concept of stability Slater claims that it is determined by the interests of the relevant
theoretical domains e.g. physics, chemistry, biology and so on, undertaking the
classification. He focuses on four different factors on the basis of which different domains
will determine stability. These are the following:

1. Level of cluster cohesiveness: Different disciplines may tolerate different degrees
   of flexibility in the clustering required by their respective kinds. Slater gives the
   example of physical kinds like electron or quark, which are supposed by those
disciplines to be perfectly clustered (conjunctive) kinds, while those of higher
biological, taxa-like families may be quite loose (Slater 2015, 403).
2. Number of properties: The number of properties that need to be co-instantiated
   together so that something is considered a kind may differ depending on the
   theoretical domain.
3. Level of ‘probabilistic entailment’ (Slater 2015, 403): Slater explains this as the
   likelihood that the instantiation of a certain sub-cluster of properties entails the
   instantiation of the whole cluster. To use the example of cluster YZT from the
   previous section, this would track the likelihood that the sub-cluster YZ would
   entail the instantiation of the whole cluster YZT.
4. Some kinds are domain relative: Slater claims that a particular kind may meet the
   requirements in one domain but fail to meet them in another. This means that
some collections of things only constitute natural kinds from the perspective of
particular sciences (Slater 2015, 403).
According to Slater, the advantages of his account compared to Boyd’s is that it evades the problems with mechanisms’ role in the HPC account and it is able to represent a more general account of natural kinds to include kinds with intrinsic essences, HPCs, causal kinds (in the sense argued by Khalidi) and non-causal kinds, that is kinds for which no causal story can be found. Slater is therefore shifting the focus from tracking causal relations of properties to tracking the stability shown by clustered properties that is the probability that the instantiation of some properties entails the instantiation of others.

In what follows I argue that Slater provides a plausible account of natural kinds, which accommodates several types of kinds that show an aptness for successful induction and explanation even in the absence of identical ‘essences’ or identical causal factors. I assess Slater’s account by discussing its relation to pluralism (e.g. the view that there are many legitimate ways of dividing the world into kinds), theory-dependence (e.g. the view that natural kinds are social constructions dependent on theoretical backgrounds) and conventionalism (e.g. the view that natural kinds depend on the interests, purposes and methodological practices of particular theoretical domains).

2.6.1 SPCs and pluralism

Slater’s account endorses pluralism about natural kinds in a similar way to Dupré, who argues that ‘there are many equally legitimate ways of dividing the world into kinds’ (Dupré 1993). Dupré rejects the idea that it is appropriate to ‘ask of an object ‘What is the natural kind to which it belongs” and claims that which natural kind a thing belongs to depends on the ‘goal underlying the intent to classify the object’ (Dupré 1993, 5). In other words the same entity could belong to more than one natural kind. Importantly for pluralism, this allows for cross-cutting kinds where objects could be members of more than one kind at a time. Those kinds are not related by any hierarchy and as such, this view rejects condition 5 as discussed in 2.2, according to which a hierarchy must be formed in such cases where the same entity could belong to more than one natural kind (Ellis 2002, 82).

According to ‘promiscuous’ realism, as it is dubbed by Dupré, there are many legitimate ways to classify the world into kinds. These ways can cross-cut each other depending on
the different purposes of the people who are doing the classifying. For example, object $x$ is a member of kind $Y$ on the basis of possessing properties $a$, $b$ and $c$. At the same time object $x$ can be a member of kind $Z$ on the basis of possessing properties $a$ and $b$; $x$ may also be a member of kind $E$ on the basis of possessing properties $a$ and $d$ or a member of kind $W$ on the basis of possessing properties $b$ and $e$.

Accepting that there can be many legitimate ways to classify the world into kinds entails that a term can be a kind in one discipline but not in another. To give an example, the term ‘jade’ refers to a kind used in jewellery production but not in chemistry. ‘Jade’ as used in jewellery production refers to both jadeite and nephrite because these minerals have similar superficial properties. However chemistry does not consider the kind jade as a chemical kind and as such only uses the classifications of ‘jadeite’ and ‘nephrite’ on the basis of their underlying microstructure. Jewellery production is governed by different interests than the ones that guide the science of chemistry. This is the case even if the object under classification has a specific underlying nature (one could call this intrinsic essence). What will determine whether the term ‘jade’, as used in jewellery production, is a natural kind term is not the proximity of the vernacular extension (the one used in jewellery production) to the scientific extension (the one used in chemistry). If one accepts pluralism, what determines whether ‘jade’ is a natural kind as used in jewellery production is the goal underlying this specific classificatory exercise.

Unless we allow for pluralism of natural kinds, debates arising between different scientific disciplines which privilege different features when determining natural kind status will always attract the charge of conventionalism. Different scientific disciplines can each privilege different features. The permissive account will not privilege one taxonomy over another as we can discover different kinds depending on which features we attend to, how we describe these features and what boundaries we draw between them.

According to this account, endorsing pluralism does not entail that all kinds are natural kinds. Although our interests determine the purposes of the classifications we create, our classifications are constrained by how the world actually is. For example, let’s assume that our goal is to make categories which ensure we find treatments against pathogenic bacteria (that is this is our classification purpose). A good way to do this is to group together all bacteria that are pathogenic to humans, and analyse their features and the way these features interact with humans. Both the intrinsic features of bacteria observed by scientists
and the way these features interact with humans are independent facts of the matter and can be determined objectively. If scientists want to make useful categories for the treatment of bacterial infections they will observe the features and behaviour of bacteria to determine which are pathogenic and which are not. On the other hand if our goal is to classify all bacterial organisms, we will use as the basis of classification either their morphological features or their phylogenetic history; we will not make categories on the basis of which bacteria are harmful to humans. Both of the above classifications could be natural kinds as long as they show an aptness for successful explanation and induction.

In a similar way, in the example of jade used above, it is an objective fact which kinds of mineral look a certain way and which can be crafted a certain way; it is not an arbitrary matter which minerals count as jade. This allows for many alternative and equally valid ways of joining individuals into groups or kinds, with different classifications serving different projects (Khalidi 2013, Dupré 2012).

2.6.2 SPCs, theory dependence and conventionalism

Slater provides a permissive account that can accommodate more natural kind candidates such as kinds with intrinsic essences, HPCs, causal kinds and non-causal kinds. According to this view, natural kindness is multiply realised: the co-instantiation of the properties that determines natural kindness is explained by different factors in different cases. Slater’s account can accommodate the types of kinds analysed by Boyd and Khalidi, and it can also accommodate kinds with intrinsic essences. What is shared by all natural kinds is an aptness for successful induction and explanation, which is a result of the fact that some properties are co-instantiated reliably.

Natural kinds are importantly mind-dependent as they are both a product of our cognitive capacity and disposition to categorise (see Barrett & Russell 2015) and also a product of theoretical domains we construct. Boyd, in a similar way, claims that the relations we discover between properties shared by the members of kinds are ‘concept, interest, and language involving’ and we discover them only within particular disciplines, that is, they are theory-dependent (Boyd 2010, 220). Boyd argues that:
The theory of natural kinds just is (nothing but) the theory of how accommodation is (sometimes) achieved between our linguistic, classificatory, and inferential practices and the causal structure of the world...... Biological taxonomists sometimes speak of the ‘erection’ of higher taxa, treating such taxa as, in a sense, human constructions. They are right – and the same thing is true of natural kinds in general. Natural kinds are social constructions: they are the workmanship of women and men. (Boyd 2010, 220)

The claim here is that naturalness is relative to and dependent on our theoretical domains. In other words, calling a kind ‘natural’ is an indication that it passes certain tests set by the relevant theoretical domain. As Boyd claims above, it appears to be the case that natural kinds cannot exist independently of linguistic and methodological practices. They are connected to our inductive and inferential practices and could not exist independently from them. However, although these categories are constructed by us, there is nothing which precludes the fact that these categories are objectively determined in the sense that presumably other creatures would also construct them if they were in the business of classifying the same entities for similar purposes.

One could argue that the above sense of ‘objectively determined’ is too weak for this account to be a realist account, if one interprets realism about natural kinds to entail the strong ontological commitment that they are real entities e.g. ‘the difference between silver and gold is not just a difference between two natural groups of thing, but is a difference between two distinct entities silver and gold’ (Bird & Tobin 2018, § 1.2). This is not a worry for Slater as he does not take the category of natural kinds to be an ontological category. Instead he argues that they can be regarded as genuine features of the world for the relevant theoretical domains because they are based on discovering stable co-instantiations of real features of the world (Slater 2015, 407). He states that the natural kind status of a property cluster:

Can be treated as a fixed objective matter when it is highly insensitive to the differences across our classificatory norms and practices. (Slater 2015, 407)

According to this view, even if natural kinds are not an ontological category they can still be treated as fixed objective matters by considering norms and practices of several different domains. This rejects the claim that natural kinds have ontological status but
preserves the epistemological value of natural kinds. Slater argues there may be some categories which play an important role in more than one discipline regardless of the difference in the norms and aims of these disciplines. He states:

Perhaps there are some clusters of properties that no matter how a discipline adjusted its norms and aims (compatible with the discipline maintaining an understanding of the natural world in view), the category that cluster described would be fit to play a robust role in the discipline. (Slater 2015, 405)

Such cases reveal the connection between this account and realism in the sense that the categories track the stable co-instantiation of real features in this world. It rejects the claim that this co-instantiation is a matter of convention.

In addition, as the SPC account also endorses pluralism, it could be argued that it is too permissive and that it lacks a way to differentiate between natural and non-natural kinds. However, this view does not entail that all kinds are natural; some kinds, such as the kind purple socks in my wardrobe, are not likely to be natural. As it happens, I have two purple socks in my wardrobe, they are not a pair, they are made from different materials, they are not the same size, they have different designs, one of them was a gift and so on. Any clustered properties associated with this group will not show stability. For example, the fact that one of the purple socks is made from cotton and is 29cm long, does not make it probable that these properties will also be co-instantiated in the other purple sock. As a result, it is not likely that these properties will be co-instantiated with stability.

Another possible objection to the SPC account is to argue that it is wholly dependent on humans and our cognitive capacities rather than on how the world actually is. According to SPC, however, whether causal or any other relations exist between properties is not something which depends on our cognitive abilities, it is supposedly a matter of fact. What depends on our interests and purposes is which of these relations we will discover and which we will not. Therefore, the natural kinds we discover are constrained by the structure, or lack thereof, of the cosmos. Patterns of relations, where these exist, constrain how the world actually is and, as a result, what kinds can be discovered by us. What maintains this relation with realism is the stability shown by the clustered properties,

---

32 However it needs to be stressed that postulating causal relations that constrain what kinds exist in the world is a separate metaphysical assumption which needs to be argued for.
regardless of any underlying causal relations. Therefore in an important sense the world adds constraints to what property clusters exist and the stability shown by the clustered properties. However, at the same time, which property cluster we will focus on depends on our classificatory purposes and interests. Haslanger for example states that:

We distinguish things by classifying them, and classification is a human activity and can be done in better or worse ways. The adequacy of a classification will depend on a variety of factors, including the extent to which the classification in question serves one’s goals and purposes, and the extent to which the goals and purposes are legitimate. (Haslanger 2012, 188)

On a weak realist interpretation of the quote above, what makes the goals and purposes legitimate is the relation between our goals and purposes and how the world actually is. According to such a weak realist view, although how we categorise the world partly and non-trivially depends on our representations and our purposes, those representations are nonetheless constrained by how the world actually is.

What is important, according to Slater’s view, is that there is no metaphysically special status guaranteed by satisfying the condition for natural kindness. Many kinds from different domains and disciplines can qualify as natural kinds on the basis of the stability portrayed by the clustered properties. As determining stability is a matter of calculating the probability of co-instantiation this account rejects any sort of metaphysical priority or fundamentality assigned by the essentialist view. This permissive view has the advantage of accommodating more types of kinds without presupposing that the clustering happens necessarily because of the possession of intrinsic essential properties, or of causal factors. At the same time, it does not reject the claim that where such relations exist, their revelation increases our knowledge and understanding of the phenomena under investigation. To conclude the SPC account accepts condition 2 (aptness for successful induction and explanation) as the only necessary (and sufficient) condition for natural kind status. All other conditions are not necessary for natural kind status.

---

33 I interpret weak realism to reject the claim that natural kinds are separate entities but accept that they are constrained by what properties exist in the world.
2.7 The SPC account as a plausible account of natural kinds

The accounts included under the epistemological view e.g. HPC, CPC and SPC agree on an important claim: The natural kind status of a kind (or property cluster) tracks the aptness of the kind for successful induction and explanation. On the other hand these accounts disagree on what explains this aptness. HPC and CPC explain this aptness on the basis of causal mechanisms or causal relations. Slater argues that this aptness is based on the stability by which the properties cluster together. In this thesis I agree with Slater’s claim that what explains the aptness is the stability of the clustered properties. It is because these properties reliably co-occur together that we are able to create successful induction following observation of individual instances.

The SPC account has certain advantages over the causal property accounts discussed in the previous section. Firstly, removing the emphasis from tracking causal relations to tracking the stability of properties allows for functional kinds to count as natural kinds. Functional kinds are defined by their functional roles rather than by the properties that realise those roles. To give an example the mental kind of pain refers to the state of being in pain regardless of what properties realise this state in different organisms e.g. in humans the state may be realised by C-fibres, in octopi by different nerve fibres and so on. Another way to describe the claim that these roles are realised by different properties is to say that functional kinds are multiply realisable. If these kinds are realised by different physical systems, it is likely that different causal factors will be responsible for the stability of the clustered properties. This way we can explain cases where other life forms exhibit emotional behaviour which is considered similar to our emotional behaviour whilst having very different physical systems realising such behaviour. In such cases if one accepts that we should individuate each psychological phenomenon on the basis of the causal mechanism or factor that underlies it, one would have to reject the claim that other life forms are capable of experiencing such phenomena.

This is related to an additional claim associated with Slater’s account that is the claim that ‘natural kindness’ is multiply realisable; this means that what is shared between all natural kinds is not some intrinsic essential property but the fact that they realise the same role e.g. they permit successful induction and explanation. This allows that the kind ‘natural kinds’ is a functional kind.
A second advantage of Slater’s account is that it allows for pluralism and for degrees of naturalness. As the concept of stability is something that can be quantified, some kinds will be more natural than others. There may be kinds where the properties always co-instantiate, e.g. the properties of quarks. On the other side there may be kinds where the properties show a degree of stability but do not always co-instantiate, e.g. the properties of biological species.

On the basis of the above, I claim that Slater provides a better account of natural kinds, one that preserves the usefulness of the concept of natural kinds without being committed to strong metaphysical claims, e.g. that natural kinds are entities or that all natural kinds track causal relations in nature. At the same time the concept of stability is something that can be objectively determined by observation. It is probable that there will be borderline cases where it will be difficult to determine whether a property cluster shows a high enough degree of stability or not and different domains may require different degrees of stability. At the same time it is possible that some property clusters show high degrees of stability in several disciplines e.g. the kind of gold is used as a category in chemistry and jewellery.

To conclude: Slater’s view provides a more plausible account of natural kinds as it is committed to weak metaphysical claims, accepts that kinds may be theory and discipline dependent, allows for degrees of naturalness and can accommodate more kinds used in scientific practice for which we can create successful induction and explanation but which do not share identical causal stories.

2.8 Summing up

In this chapter I gave a broad description of two different views on natural kinds. According to the metaphysical view, natural kinds are metaphysically fundamental in the sense that they correspond to pre-established natural ‘joints’. As I mentioned in the previous section the metaphysical view could be associated with the strong ontological commitment that each natural kind refers to an individual entity. However one can still hold the metaphysical view that natural kinds are fundamental without also accepting that this is due to the fact that they are individual entities. Instead one can hold that this fundamentality is related to possessing intrinsic essences which are shared by all members of the kind. Most of my arguments in this thesis are directed against the latter view.
kind members and what explains the aptness shown for induction and explanation by the category. These intrinsic essences are supposedly discovered purely *a posteriori*: they cannot be stipulated in advance as this jeopardises the category's independent (of the human mind) existence. The metaphysical view takes several conditions to attach to natural kinds necessarily, such as sharing intrinsic properties, forming hierarchies, forming kinds, etc. I argued that it is not clear why these conditions should be thought to attach to natural kinds, or whether these are postulated *a priori* or whether they are discovered exclusively *a posteriori*. According to the metaphysical view, there are very few natural kinds, mostly matching divisions used in the sciences of physics and chemistry (Ellis 2001; 2002).

I discussed two criticisms of the metaphysical view. The first one relates to the claim that intrinsic essences are discovered exclusively *a posteriori*. As LaPorte claims, we have cases in chemistry where it is shown that such intrinsic essences are not ‘discovered’ but are stipulated. In addition I argued that identifying natural kindness with the possession of intrinsic essences, unnecessarily restricts the number of natural kinds to a few kinds in physics and chemistry. By contrast, I argued that the epistemological view can be applied to a larger number of groupings in nature. It does not consider natural kinds to be metaphysically fundamental entities. Instead, according to the epistemological view, natural kinds are useful ways to categorise objects such that they facilitate induction and explanation. It is primarily this role that they play which makes them fundamentally important for science. Although I call this view epistemological to distinguish it from the metaphysical view, it is associated with the weak metaphysical claim that our successful classifications track some properties which can be determined objectively e.g. causal relations according to causal property cluster accounts or the property of showing stability according to the stable property cluster account.

I discussed Boyd’s account, which can be applied to kinds both from the biological and social fields. I also discussed Khalidi’s purely causal account, which focuses on the existence of causal relations as what explains the aptness for induction and explanation. I accepted Craver’s insight that appealing to mechanisms invites conventionalism in the process of classification, however, I also accept his claim this is not a serious threat to the usefulness of the concept of natural kinds or to property cluster accounts’ realist credentials.35 I then briefly discussed Slater’s (2015) more relaxed version of the

---

35 I will argue for this claim in chapters 3 where I analyse social kinds and chapter 4 where I focus on the kind of emotion.
epistemological view, which can be applied to groupings in the absence of causal relations. All these epistemological accounts preserve the usefulness of natural kinds and consider them epistemically superior. There are several questions relating to the concept of stability used by Slater; however this account is a good basis for a more permissive view of natural kinds.

The accounts falling under the epistemological view are compatible with pluralism of the type advocated by Dupré (1993) and they are also compatible with the claim that natural kindness comes in degrees (Slater 2015). Accepting pluralism and degrees of naturalness implies rejecting some of the characteristics usually associated with natural kinds such as having clear and distinct boundaries and hierarchies between natural kinds (Bird and Tobin 2018, §1).

Natural kind status is not restricted to the kinds of entities usually classified under microphysical or chemical domains. They can be discovered anywhere in nature wherever and whenever clusters of properties show stability. This account does not reject a distinction between natural and non-natural kinds. Instead this distinction becomes one of degree, e.g. gold may be more natural than tiger, which in turn may be more natural than book. Naturalness tracks the stability with which the properties cluster together. This stability is not restricted to the microphysical or chemical sphere. It can appear in many different domains, scientific and non-scientific. This includes both the social and psychological domains, which have traditionally been claimed to be areas where natural kinds cannot exist. In the next chapter I argue that by accepting the epistemological view and more specifically Slater’s SPC account, one can successfully claim that some social or psychological kinds can be considered natural kinds.
Chapter 3- Socially Constructed Natural kinds

In this thesis I put forward a claim that emotion is a social kind. Social kinds are typically used as examples of kinds that cannot be considered natural. In this chapter I discuss three reasons used to support the claim that social kinds cannot be considered natural; first, because they result from heterogeneous construction; second, because they are interactive and third, because they change in accordance with normative factors. I argue that the account of natural kinds I endorsed in chapter 2 can provide replies to such objections such that some social kinds can be considered natural kinds.

3.1 Social kinds vs. natural kinds

According to the traditional view on natural kinds (which I called the metaphysical view in chapter 2), natural kinds are kinds that are wholly independent from us, our representations of the world and our classificatory purposes. According to the metaphysical view any kind which depends on human societies and their practices is not natural but ‘social’; social kinds are opposed to natural kinds by definition. It seems that according to the metaphysical view, ‘social kinds’ and ‘socially constructed kinds’ refer to the same kinds. They are social because they depend on us and our social practices rather than being wholly independent of them. I will call this the ‘opposite-by-definition’ view.

On the other hand, according to an alternative view, human societies and their practices are parts of the natural world; there is no contradiction between the terms ‘natural’ and ‘social’. The social sphere can be considered as part of the natural world; human societies are parts of nature. They do not exist outside the sphere of the natural. In a similar way Haslanger claims that:

"There is an important sense, moreover, in which the natural world is all there is. Our activities, our minds and their contents, our language and its meaning, our interactions, are all natural (or supervene on the natural)...Rather than distinguishing the natural and the social, the social should be understood as part of the natural. (Haslanger 2012, 213)"

According to this interpretation, the social is considered part of the natural. I will call this the ‘part-of-nature’ interpretation. There is no contrast between the natural and the social. The contrast only exists between the social and the non-social within the natural.
The epistemological view is closer to the part-of-nature view as it does not consider the terms ‘social’ and ‘natural’ to be opposite by definition. Natural kinds are by definition socially constructed, in the sense that they partly depend on the practices, methods and classification purposes of individual theoretical domains. According to the epistemological view, there is no embedded contradiction between natural and social kinds. As Boyd mentions, ‘natural kinds are the social constructions of women and men’ (Boyd 2010, 220). In addition, according to the epistemological view, some of these natural kinds are social kinds. Social kinds, unlike other kinds such as chemical kinds, relate to social phenomena, phenomena whose analysis and explanation requires investigation of the social domain. Therefore according to the epistemological view there is a difference between the terms ‘socially constructed’ and ‘social’.

In this chapter I argue that several objections to the claim that social kinds can be considered natural kinds are based on accepting essentialist metaphysical assumptions. If these assumptions are not accepted, these objections have no force. The aim of this chapter is to uncover these assumptions so that their connection to essentialism is made obvious. I conclude that the relaxation of the conditions under which a kind can be considered natural, as required by the epistemological view and more specifically by Slater’s view, has the consequence that many kinds which traditionally have been considered as non-natural, such as some social or psychological kinds, are now candidates for natural kind status.

Most of the objections to the natural kind status of social kinds involve a claim of heterogeneous construction, e.g. the claim that members of social kinds have the properties they do because of different causal factors. However I argue that once one accepts that discovering natural kinds is the process of tracking the aptness shown for induction and explanation, the causal stories associated with the property clusters are complementary to but not determinants of the natural kind status of the categories under investigation. Hence heterogeneous construction turns out to be no barrier to natural kindhood.

In the first part of this chapter I discuss three common objections to the claim that social kinds can be considered natural kinds. In the second part I provide replies to those objections and show how they are connected to the different views about the relation between the social and the natural. I focus on the categories of gender, e.g. woman or man,
and emotion, which are used in the literature as examples of socially constructed social kinds which do not qualify as natural kinds (Griffiths 2004, 901). This discussion will show that social kinds can be considered natural kinds on the basis of the stability portrayed by the cluster properties regardless of whether these clusters a) are associated with different causal origins (heterogeneity), or b) have changing properties as a result of interacting with the environment (interactivity and normativity).

3.2 Objections to the claim that social kinds can be considered natural

Philosophers who accept the opposite-by-definition view accept that social kinds cannot be considered natural kinds (Ellis 2001). According to such views natural kinds are limited to a few microphysical or chemical kinds (Ellis 2002, 154). On the other hand other philosophers, such as Griffiths, accept the part-of-nature view and claim that some social kinds can be considered natural kinds whilst others cannot (2004). Although these views make different claims on the natural kind status of social kinds such as gender categories or emotion, they raise similar objections to the natural kind status of social kinds. In this chapter I focus on objections based on an examination of the relations between the property clusters and the environment. In the following chapters I argue that it is precisely the relation between the property clusters and the environment that justifies the claim that these kinds are social rather than biological or neurobiological.

As mentioned in chapter 2, one way to differentiate between natural kinds is to categorise them on the basis of causal factors. This way of distinguishing natural kinds in terms of examining causal relations gives rise to three common objections to the claim that social kinds can be considered natural; these relate to the features of heterogeneity, interactivity and normativity and are explained below:

3.2.1 Heterogeneity

Some philosophers argue that social kinds such as some psychological kinds do not arise as a result of the same kinds of causal factor as do genuine natural kinds (Hacking 1992). In particular, social and psychological kinds are too heterogeneous to count as natural

---

36 According to this view this claim would also apply to socially constructed kinds as these are considered identical.
kinds (Griffiths 1997, 229). For example, according to this objection psychological phenomena, such as an instance of ‘fear of spiders’ and an instance of ‘anger at economic austerity measures’, cannot be members of the same natural kind because they have ‘different phylogenies, different adaptive functions, different neuroscience and different roles in human psychology’ (Griffiths 1997, 16). Instead, they refer to separate psychological kinds; these separate kinds may or may not be natural kinds but the overall category of emotion which includes them both is not.

According to this view, a group of phenomena would be members of the same natural kind only if they have the same or sufficiently similar causes. Griffiths claims that ‘the purpose of categorisation in the sciences is to group together things which resemble one another in many different ways because of some underlying similarity-generating mechanism’ (Griffiths 1997, 16). If the criterion on the basis of which we define kinds is this ‘underlying similarity-generating mechanism’, objects with different mechanisms cannot be considered members of the same natural kind; they are heterogeneous. Griffiths claims that:

> The extension of a concept is determined by our best current theory of the causal homeostatic mechanism that guarantees the projectibility of the category to which it refers… If there is no such mechanism, then the concept fails to refer to a natural kind and should be discarded for the purposes of induction and explanation. (1997, 246)

According to Griffiths, for fear and anger to be members of the same natural kind or for all instances of what in everyday language we call ‘fear’ to be members of the same natural kind of fear, the mechanisms that underlie these phenomena need to be of the same kind by sharing the same ‘causal homeostatic mechanism’. Anger at economic austerity measures seems to include a cognitive element which originates from a psychological mechanism which is different to the one associated with fear of spiders. According to this view, the mechanisms underlying all emotional phenomena must be of the same kind so that the category of emotion is considered a natural kind.

---

37 As in previous chapters, I use italics to refer to kinds of individual emotion such as fear to differentiate them from instances of fear.
3.2.2 Interactivity

One of the difficulties in determining the intension and extension of social kind terms such as gender or emotion concepts is the fact that, as Haslanger argues, in social kinds our concepts and our social practices can be deeply intertwined:

The main point to note here is that our classificatory schemes, at least in social contexts, may do more than just map pre-existing groups of individuals; rather our attributions have the power to both establish and reinforce groupings which may eventually come to ‘fit’ the classifications. (Haslanger 2012, 88)

Haslanger claims that gender categories are kinds constructed as a result of an interplay between social practices and individuals. She argues that:

Gender classifications occur within a complex matrix of institutions and practices, and being classified as a woman, or a man, or a different sex/gender, or not, has a profound effect on an individual. Such classification will have a material effect on her social position as well as affect her experience and self-understanding. (Haslanger 2017, 167)

This difficulty in determining the intension and extension of the terms is related to Hacking’s (1999, 108) suggestion regarding ‘looping effects’ in categorisation, where the individual members are influenced by being categorised as one thing or another and their behaviour changes either to conform towards the categorisation or to act against it. Hacking argues that persons classified in a certain way may come to change in response to the labels placed upon them (1999, 108). This looping effect destabilizes human kinds and our knowledge of them. For example, the referent of the term ‘woman’ seems to be evolving both as a result of external forces such as the way of classification set by society and as a result of individuals choosing to either conform to or avoid conforming to the criteria for membership. Hacking calls such kinds ‘interactive kinds’; there is interaction between the way one is classified and the person in the sense that the person either conforms to or changes their behaviour and characteristics after becoming aware of how they are being classified (1999, 32). There are no interactions of the same type in the natural sciences in the sense that the quark does not care if it is categorised as quark. It is
not aware of its classification as one thing or another. Hacking explains this process as follows:

People of these kinds can become aware that they are classified as such. They can make tacit or even explicit choices, adapt or adopt ways of living so as to fit or get away from the very classification that may be applied to them…The result may be particularly strong interactions. What was known about people of a kind may become false because people of that kind have changed in virtue of what they believe about themselves. (Hacking 1999, 34)

Our practices of thinking about, describing, and differentially treating persons can give rise to the very kinds that those practices describe and refer to in the social domain. For example, the referent of the term ‘woman’ depends on the society which constructs the kind. Old and new theories about the referent of the term ‘woman’ will be talking about different things i.e. they will not co-refer. According to Hacking, similar issues attach to other social kinds such as emotion. He claims that emotions vary between cultures in their physiological expressions and that how we describe emotions affects how emotions are experienced (Hacking 1999, 18).

As discussed in the previous chapter, several social or psychological kinds are made via convention, legislation or decree (Hacking 1999, 10; Khalidi 2015, 1385). Hacking gives the category of woman refugee as an example of a socially constructed kind. He argues that the idea of the woman refugee is constructed; this way of classifying people is the product of social events, legislation, social workers, immigrant groups, activists, lawyers, and of the activities of the woman involved. This kind of person, as a specific kind of person, is socially constructed (Hacking 1999, 10). However, members of such kinds can come to share common properties as a result of the classification. This is because being classified as x acts as a causal factor for either exhibiting properties associated with x or for rejecting properties associated with x. There is an ongoing interaction between the fact that these members are being classified as x and the members exhibiting the properties characteristic of x.

Hacking uses MPD (Multiple Personality Disorder currently known as Dissociative Identity Disorder) as an alternative example of such an interactive kind (1995). Hacking claims that psychiatrists form the classification of MPD on the basis of patients exhibiting alternating personalities (Hacking 1995). Knowledge of this classification and the
properties it tracks enter the public domain. Some of the people diagnosed with MPD become aware of the properties associated with this disorder. They then start exhibiting new symptoms such as exhibiting animal personalities. Therefore knowledge of the classification by some people diagnosed with MPD can result in a new property associated with the kind and a new kind of people that is people who exhibit animal personalities.

Hacking claims that some of the people who are classified/diagnosed as members of the kind MPD and treated accordingly, come to identify with the kind and acquire new properties which they did not possess in advance of the classification and treatment, or stop exhibiting properties associated with the kind. However, both these changes have an effect on properties exhibited by kind members and as such, ultimately influence the properties associated with MPD. Before the classification became common knowledge in the society most kind members did not exhibit the property of ‘exhibiting an animal personality’. However after this property becomes known as a property associated with this disorder some patients who become aware of this property but did not exhibit it in the past, start to exhibit it. This constant interaction, according to Hacking, makes it difficult to determine what the properties of the kind are and how we will determine membership. Hacking concludes that:

The targets of the natural sciences are stationary. Because of looping effects, the targets of the social sciences are on the move. (Hacking 1999, 108)

The interactivity objection runs as follows: ‘interactive kinds wholly or partly depend on human thoughts and actions in ways that kinds such as gold, or tiger supposedly do not’. According to this objection, the clustered properties characteristic of such kinds are constantly changing in a way that makes it difficult to pin down the characteristic properties of the kind and determine membership. Members of social kinds such as gender categories are influenced by the fact that they are classified one way or another by either conforming to the classification or by objecting to it (Hacking 1999, 10). Such interactive clusters cannot be considered natural kinds because it is impossible to identify the clustered properties as these keep changing. Hacking states that human kinds are ‘different from what philosophers call natural kinds because they interact with the very beings to which they apply’ (1991, 258).

---

38 I borrow this example from Tekin (2014) who offers a good critique of Hacking’s position.
A first thing to note here is that Hacking can be interpreted as claiming that all human kinds are interactive kinds. According to Khalidi we should interpret Hacking as claiming that all interactive kinds are human kinds instead (2013, 147). Although this question is interesting, for the purposes of this thesis I focus on Hacking’s claim that the feature of interactivity is incompatible with natural kinds. As Hacking claims above, human kinds are different to natural kinds because of this feature of interactivity. I provide a reply to this objection in 3.3.2 below.

### 3.2.3 Normativity

Some social or psychological kinds such as *child abuse* or *emotion* allow normative considerations to influence what properties should and should not be included in the property clusters (Griffiths 2004, 909; Hacking 1992, 188). According to this view the properties that will be included in the cluster *child abuse* do not only reflect what exists in nature but also the changing norms of society. What is considered ‘child abuse’ today, was not considered ‘child abuse’ 50 years ago. This change in the diagnostic criteria does not necessarily reflect the epistemic project of increasing the inductive and explanatory power of the category (Griffiths 2004, 906); instead, according to this view, it represents a new normative model of the conception of abusing a child (Hacking 1991, 253).

In a similar way, Griffiths claims that the category of *emotion* has a normative dimension in allowing for evaluating properties to be included in the property cluster which influence the inductive and explanatory power of the category (Griffiths 2004, 909). For example, normative considerations attach to when someone should show emotion, the degree of such expression, and so on. To give a concrete example, when Georgia is experiencing *sadness* because her team lost the championship in the finals, there are norms dictating how long this emotion should be experienced for (e.g. two months would be too long), what expressive actions should be appropriate (e.g. falling into a depressive state or resorting to physical violence such as pulling out one’s hair would not be acceptable) and so on. Griffiths claims that these normative considerations influence the classification of emotions as they involve two different and conflicting projects; the first project relates to the description of emotional episodes whilst the second one has a prescriptive role as it dictates how and whether the emotion will be experienced (2004, 908). This influence invites convention into the process of classification as we no longer objectively determine which properties cluster together; instead societal norms determine which properties will
cluster together. In the example used above, instead of simply observing the duration, degree and expressive action associated with the emotion of sadness, societal norms dictate how this emotion should be experienced; they have a prescriptive role that interferes with simple observation of instances.

Because of these features influencing emotional phenomena, Griffiths takes an eliminativist position about ‘emotion’. The term ‘emotion’ and the terms we use for individual emotions such as ‘fear’ are examples of ‘partial reference’ (Griffiths 2004, 901). Griffiths provides the following example to explain what he means by ‘partial reference’:

The term ‘jade’ is the standard example of partial reference. ‘Jade’ may be either of two different stones, jadeite or nephrite, and the term ‘jade’ partially refers to each of these two kinds of stone. It follows from this fact that for the purposes of geology or chemistry, jade cannot be treated as a single kind of thing. (Griffiths 2004, 901)

Griffiths claims that the term ‘jade’ refers to two different minerals. In geology and chemistry there is no category of jade, precisely because the term refers to two different minerals. He therefore argues that the same should apply to the term ‘emotion’ if the term refers to different phenomena.

To give an example of a kind that arguably includes all three features discussed – namely heterogeneity, interactivity and normativity – consider the psychiatric kind of ADHD (Attention Deficit Hyperactivity Disorder). It is a matter of considerable debate whether ADHD refers to a single psychiatric kind or whether it refers to two different psychiatric kinds that are caused by distinct factors, where the latter would rule ADHD out from counting as a natural kind according to the heterogeneity objection (See Furman 2008). In addition it is argued that it is not clear that the properties included in the cluster of ADHD are not a result of several different causal factors such as looping effects (the interactivity objection; see Kendler, Zachar & Craver 2011). So according to the interactivity objection, discovering the symptoms of ADHD should strictly be a descriptive process; psychiatrists would observe patient’ symptoms and describe the property cluster associated with ADHD. However looping effects interfere with the

---

39 For a discussion of ADHD as a candidate for natural kind status see Khalidi (2013, 191). Khalidi provides a reply to all objections to claim that a specific type of ADHD may be a good candidate for natural kind status on the basis of our current knowledge of this disorder.
descriptive project in that the patients of ADHD may start exhibiting symptoms which they did not exhibit before as a result of being diagnosed as ADHD patients. Finally it is claimed that the kind is too conventional by allowing normative considerations such as the interests of pharmaceutical companies in its definition and the description of symptoms and treatment (the normativity objection; see Furman 2008; Kendler, Zachar & Craver 2011).

In what follows I focus on the metaphysics of social kinds to reply to the three objections described above. I argue that the objections arise from holding assumptions relating to the metaphysical view of natural kinds. However, if one accepts the epistemological view, and more specifically Slater’s account, one does not need to subscribe to any metaphysical assumptions relating to essentialism about kinds.

As this thesis focuses on scientific categories and more specifically categories from the social and psychological domains, I argue that whether the candidate kinds show an aptness for successful induction and explanation is a matter which can be determined objectively within each scientific domain. An important assumption that I will be maintaining throughout this chapter is that if one scientific domain determines that a category is not a natural kind for the purposes of that domain, this does not entail that the category is not a natural kind in any scientific domain.40 Throughout the discussion of these objections I focus on the candidate kinds of gender categories and emotion and discuss the circumstances under which these kinds could be considered natural kinds according to the account I endorsed in chapter 2.

---

40 This point will become more relevant in the following chapter where I focus on the natural kind status of emotion.
3.3 Replies to objections

3.3.1 On the heterogeneity of social kinds

According to the objection from heterogeneity, members of some kinds exhibit the properties by virtue of which they are members of those kinds as a result of different factors or mechanisms; hence such kinds cannot be natural kinds. According to this view there is a difference in the generating mechanisms and on that basis members of some kinds cannot be considered natural. The objection from heterogeneity can be related to reductionist projects in the explanation of social or psychological phenomena. According to such projects in order for social or psychological categories such as gender categories or emotion to be candidates for natural kind status they need to be reduced to some type of biological or neurophysiological category. According to this view, an absence of a single biological or neurophysiological mechanism responsible for the clustering of the properties in the kind emotion entails that this kind cannot be considered a natural kind. For example, according to Griffiths, the existence of several different neurophysiological mechanisms underlying instances of emotion entails that we are actually dealing with different kinds (1997, 246).

There are several possible interpretations of the objection from heterogeneity. A first interpretation is that difference in the causal factors entails difference in the property clusters. For example, Griffiths’ claims that the category of emotion includes heterogeneous phenomena generated by different biological mechanisms. This explains the difference exhibited by the neurobiological properties between instances of emotional phenomena. I agree with Griffiths that the causal story can provide additional information on the nature of emotional phenomena and can perhaps explain why we cannot make successful inductions and explanations on the basis of the neurobiological properties of emotion. However to agree with Griffiths’ claim that emotion is not a legitimate object of scientific investigation in neurophysiology, one needs to presuppose that emotion is a neurophysiological property cluster. It is this additional assumption that I reject. I focus

---

41 As mentioned in the previous chapter, the term ‘mechanism’ can be defined in different ways depending on the phenomena under investigation. Here I use it to mean ‘factor’ or any type of causal origin.
on this issue in the next chapter. In what follows I focus on the interpretation which takes similarity of causal factors to be necessary for natural kind status.

According to the latter interpretation of the heterogeneity objection, we should define a kind on the basis of the causal mechanism that is responsible for the kind. Different causal factors entail different kinds. To give an example, we could consider kind \( \kappa \) defined as ‘whatever is caused by mechanism \( y \)’. In the example of the emotion of fear used above, the kind of fear would be defined as ‘whatever is caused by the activation of mechanism \( y \)’. As mentioned in the previous chapter, I reject such views on the basis of the fact that similarity of causal mechanisms does not appear to be necessary for natural kind status. In the remaining of this chapter I provide more detail for this claim.

As mentioned in chapter 2, what we consider as the causal ‘mechanism’ seems to depend on the specific phenomenon under investigation (Craver 2009). Depending on which property cluster we focus on, we can define its underlying mechanism in a different way. This is especially so in relation to psychological kinds where distinguishing brain mechanisms is extremely difficult and invites conventionalism into the classification process (Craver 2009). Therefore according to Craver, we lack an objective way to define causal mechanisms for psychological phenomena. Secondly, as Slater claims, what guarantees successful induction and explanation is not the existence of a causal mechanism but the stability portrayed by the clustered properties (2015). If the properties cluster with stability, we will be able to make successful induction and explanation for the clustered properties. This should be the case regardless of whether there exist the same or sufficiently similar causes for the properties in the cluster.

Slater’s point is demonstrated in cases where members of the same property cluster do not have the same or similar causes, but the properties within the cluster exhibit stability. Consider the case of the element of Plutonium, \( Pu \). Only traces of this element are found existing in nature independently of humans (IEER, 2005). Apart from those traces, all other samples of Plutonium are created by humans synthetically in the lab. The samples of Plutonium created by humans and the samples of Plutonium existing independently in nature have identical physical and chemical properties. However, they do not have the same or similar causes. This example shows that it is possible that property clusters co-instantiate reliably without necessarily arising out of the same causal factors. If this
phenomenon is observed in chemical elements which are not as complex as mental phenomena such as emotion, it is likely that it will also be observed in more complex kinds where there may be multiple causal factors involved. Therefore heterogeneity in the causal factors should not imply heterogeneity in the kind itself.

This consideration supports the claim that being caused by different factors does not entail that the property cluster will not show any stability. In the case of simple kinds such as the kind *gold*, it seems uncontroversial to look for a single causal factor, for example atomic number, considered as responsible for the clustering of the properties. In cases of complex kinds such as *tiger* or *emotion*, it is likely that there will be more than one causal factor responsible for the clustering of the properties. If we add to this that it is not a requirement that the causal factors are identical, this entails that the causal story behind the clustering may be a complex one rather than one that resembles the case of *gold*. If we wish to provide an accurate causal story of complex phenomena such as gender categories and *emotion*, we need to resist attempts to reduce these phenomena to simpler clusters as this distorts reality. Mallon, in a similar way, claims that with respect to social kinds there are several alternative explanations that can be provided for their existence (2016, 160). He states:

Metaphysically moderate constructionism insists that some things exist or have the character that they do because of mechanisms that are part of human minds such as mental states, cultural information, human actions and artefacts. They are competitors for the explanation of categories such as race, sex, gender, mental illness, the emotions etc. They oppose particular accounts of these categories that explain human difference by appeal to ‘natural’ i.e. biological, neurological or genetic facts. They offer alternative causal and constitutive explanations. (Mallon 2016, 160)

According to Mallon, phenomena such as gender categories or *emotion* cannot be explained by appeal to biological, or neurological facts. Their property clusters include features from several different domains such as psychology, neuroscience, sociology etc. If the social phenomena under investigation have features that are being studied by different disciplines, it is likely that the explanations one can provide for the clustering of the properties will be as complex and multi-dimensional as the phenomena under
investigation. However, any attempt to reduce the phenomena to the features studied by the disciplines of biology, physics or chemistry distorts their nature and should be resisted.

At the same time, accepting that some social kinds are natural kinds does not entail that all social kinds or socially constructed kinds are natural. Arbitrary collections of things which do not show any significant aptness for induction and explanation are not natural kinds. To give an example, the grouping of ‘cool people’ is unlikely to show an aptness for induction and explanation as what one considers as ‘the properties of being cool’ seems to depend on the individual who is doing the evaluating. In such a case the same property could be determined as ‘cool’ when it is exhibited by one person, and not ‘cool’ when it is exhibited by another even where the same person is evaluating the property of ‘coolness’ (see Haslanger 2000, 92).42

The objection from heterogeneity is associated with Griffiths’ claim that the term ‘emotion’ refers to different mental phenomena and, as such, partially refers. For this claim to be true one would have to agree with Griffiths that the term ‘emotion’ or the terms of individual emotions such as ‘fear’ refer to different mental phenomena. However, whether these terms refer to different phenomena depends on how one defines the phenomena themselves. If one chooses to define emotion as exclusively a basic emotion, that is as an automated response analysable in terms of neurobiological properties, which is Ekman’s (Ekman & Cordaro 2011) position, one would have to agree with Griffiths that the phenomena referred to by these terms are different and, as such, that the terms partially refer. However, there are alternative ways to define emotion which do not take it to be a strictly neurobiological phenomenon.43 If one defines emotion at a more abstract level, for example as a social role which includes several physiological, psychological, social and neurobiological properties and which cannot be reduced to any of those categories whilst retaining its nature, the claim relating to partial reference does not apply. For example we could consider fear as a social role which is realised by different physiological and neurobiological properties. In such a case difference in the neurobiological properties, or any underlying generating mechanism between an instance

---

42 Haslanger uses such a group as an example of what she calls ‘strong pragmatic construction’ where the classification is not capturing real facts and distinctions but there is an illusion that it is capturing intrinsic facts about individuals.

43 I analyse such definitions in chapter 4 of this thesis where I argue that emotion cannot be reduced to strictly neurobiological properties.
of fear and an instance of anger would not automatically entail that these instances are not members of the same stable property cluster. What it entails, is that such instances cannot be members of the same neurobiological kind which exclusively tracks sameness of neurobiological properties. To relate this to the example of jade used by Griffiths, it may be the case that jade is not a chemical kind. This fact does not imply that jade is not a natural kind. In the same way, emotion may not be a neurobiological kind. But this does not entail that it is not a natural kind.

The partial reference claim seems to rely on the assumption that the physical/neurobiological bases underlying each instance of emotion differ to such an extent that instances of different emotions do not fall under a common natural kind, emotion. This fact would support the conclusion that these instances belong to different natural kinds if one identified emotion on the basis of having a specific neurobiological underlying nature, for example emotion is ‘neurobiological state e’. However, if emotion is identified with a more abstract entity which is multiply realised, and as such, is not identical to ‘neurobiological state e’ then the simple fact that the neurobiological component between two different instances of emotion such as fear and anger, are different should not entail, on its own that these two instances are not members of the same stable property cluster.

3.3.2 On the interactivity of social kinds

The interactivity objection is based on the claim that some properties within the cluster change because members of the kind are influenced by the classification process e.g. members of the kind woman either change their behaviour to conform or reject some of the properties associated with this cluster. This objection should pose no real issue given the epistemological account of natural kinds I endorsed in chapter 2, which is based on tracking the stability of the clustered properties. As with other scientific concepts, our best scientific theories will investigate and determine whether terms such as ‘woman’ or ‘man’ refer to SPCs.

To establish whether a kind has objectively determined common features (whether it qualifies as a SPC) we need to observe paradigm cases to establish what, if anything, is shared between kind members. The objectivity feature partly relates to the direction of fit
between our representation of the world and the world itself. If there is no objectively determined commonality between members of gender categories, such kinds will not show an aptness for induction and explanation. According to social constructionist accounts of gender categories, it could be argued that the categories originally constructed by us come to exist as property clusters on the basis of members interacting with the standards set for the kind members. In the same sense chartered accountants have to conform to certain rules in order to remain members of the kind. If one does not endorse the assumption that human activities and practices cannot cause or maintain stable property clusters, one will determine whether kinds such as chartered accountants would be considered natural on the basis of examining paradigm cases. Paradigm cases need to be examined to determine the stability of the property clusters and the aptness shown by the categories for successful induction and explanation regardless of whether the kind shows features of interactivity or heterogeneity.

A key consideration relating to the objection from interactivity attaches to the fact that the clustered properties change over time. As Hacking mentioned above, the target of social kinds seems to be on the move rather than remaining unchanged. However, anyone who accepts tiger as a natural kind has to accept that this property cluster is evolving. As Mallon argues:

Mere change or evolution in the world does not undermine knowledge of it, so long as the rate of change is less than the rate at which our theories improve themselves….The test of whether a particular social role amounts to an important kind is whether making reference to it is useful in epistemic projects like explanation, prediction and intervention. (Mallon 2016, 167)

According to the epistemological view what determines whether a kind is natural is the aptness for induction and explanation. As Mallon claims above, the fact that the property cluster changes over time is not enough on its own to render the kind non-natural. If the change happens at such speed or frequency that there is no stability portrayed by the clustered properties, the kind will not show an aptness for induction and explanation. Therefore, change should not threaten the stability of gender categories or the kind emotion in a similar way that change does not threaten the reality or stability of the kind tiger. If change is an issue for interactive kinds, then it necessarily creates issues for anyone who
accepts the view that species are natural kinds. If the above considerations are correct, a category may be a natural kind even when it is an interactive kind in the sense suggested by Hacking.

Mallon explains this interactive process and claims that some social kinds such as gender categories, or *race*, or *dissociative identity disorder* may be constructed by us and our representations. Where this information is generated by a credible source it can create common knowledge about a social role which operates as a stable property cluster (2016, 93). Mallon states that:

Some such social roles, in turn, become entrenched, producing a range of effects that further differentiate putative members of the role....They may begin as somewhat narrow property-cluster kinds centred around a community’s labelling practices, and over time grow into more and more substantial property clusters.....As these property clusters grow in significance, reference to them may become increasingly important to understanding the social world. (Mallon 2016, 93)

Therefore as Mallon claims above, regardless of how a kind was constructed and regardless of the fact that the property cluster undergoes change, it can still allow for successful induction and explanation on the condition that the change can be identified.

In the previous chapter I endorsed Slater’s account, according to which, the degree of stability required for natural kind status is different depending on each discipline. This entails that different degrees of stability will be acceptable in different disciplines. So it is likely that social sciences such as economics are disciplines where induction and explanation are more difficult because the property clusters tend to be less stable. However, they are considered stable enough to be useful categories (in the sense of being apt for successful induction and explanation) in these disciplines. What this entails is that ‘looping effects’ as described by Hacking need not be considered as causes of permanent instability such that no stability is shown by the property clusters affected by them. Arguably these property clusters have an additional level of complication compared to biological categories e.g. *tiger* in that their properties are less stable. For example the kind woman may show less stability than the kind tiger. However as mentioned in chapter 2
the account I endorse allows for degrees of ‘natural kindness’ that is of the aptness shown for successful generalisation and induction. As such, according to my view social kinds with looping effects can still be considered natural kinds.

3.3.3 On the normativity of social kinds

Considerations attaching to normative projects, such as eradicating gender inequality or exhibiting appropriate emotional behaviour, give rise to a separate objection against the claim that these social kinds can be considered natural kinds. Griffiths, for example, claims that the category of emotion is simultaneously an investigative and a normative kind (Griffiths 2004). According to Griffiths, the category of emotion is pulled by these competing features in two different directions. The investigative project is aimed at describing the nature of emotional phenomena as they are manifested whereas the normative project is aimed that manipulating the same phenomena in accordance with our norms and practices. Griffiths explains this by describing the change in the definition of ‘child abuse’. He states that:

The change from viewing a pattern of childcare as normal to viewing it as abusive need not reflect an epistemic project, such as maximizing the predictive power of child abuse as a diagnostic category in psychiatry. The change can equally well represent the spread of a new normative model of the relationship between parent and child, or a change in the relative value placed on various traits of the older child or adult, such as placing a higher value on personal fulfilment and a lower value on social conformity. (2004, 908)

The feature of normativity is intimately related to the feature of interactivity which was discussed in the previous section. Norms and practices influence the properties that cluster together so that these constantly change to match changing societal attitudes to them. To give an example, society influences which type of emotion should be exhibited in a given situation, which object it should be directed towards and to what degree it should be exhibited. For example, during the romantic period (1800-1850) the emotion of anger was not considered appropriate for women as they were considered to be weak and needing of protection. There were hardly any cases where it would be appropriate for
a woman to experience or express anger. According to Griffiths this interferes with the project of investigating the nature of emotional phenomena. According to this view the nature of emotion is influenced and is constantly being altered by societal norms attaching to the expression and experience of emotion.

According to this objection, discovering which properties cluster together is not exclusively an investigative process. We not only discover which properties cluster together but we dictate which properties will or should cluster together. Griffiths claims that this is problematic when we are dealing with scientific categories as these are supposed to be discovered in rather than enforced on nature. A reply to this objection runs as follows: The normative forces that Griffiths refers to may be considered as external factors which are responsible for some of the clustering of the properties associated with emotion. It may be the case that, as we investigate the nature of emotion, we discover that normative forces influence the clustered properties. On its own this discovery cannot pre-determine that emotion will not show an aptness for induction and explanation. Normative forces can be considered as another type of external force which interacts with the property cluster. For example envy and pride were included in the ‘seven deadly sins’ in Christian doctrine. All instances of these emotions were considered inappropriate and believers of the Christian faith had to abstain from them. Religious norms influenced how people ought to feel or ought not to feel. Khalidi in a similar way states that in some cases convention is entangled with causality:

Some natural kinds also come into existence as a result of human intervention and they can interact in various ways with our thoughts and actions….Normativity is by no means a feature of all social kinds, and when it is, it can be detected. I conclude that categories in the biological and social sciences are not fundamentally different from those in the natural sciences and that biological and social kinds can be natural kinds as well. (Khalidi 2013, xiii)

As Khalidi claims above, normative forces can be considered as another causal factor influencing the property cluster. These forces can be detected and analysed in the same way as other causal factors. Their existence does not entail that the kind will not be considered a natural kind. Unless one holds additional assumptions which either exclude
human activity from being considered a natural causal factor or which require the existence of unchanging necessary essences, one need not accept that normative, interactive kinds cannot be considered natural. Therefore, unless one already assumes that human causal factors are by definition non-natural, this fact alone does not predetermine that the property cluster will not exhibit the stability required. To agree with Griffiths one would have to exclude human activities, practices and purposes from the sphere of the natural causes (what I called the ‘opposed by definition’ view). Therefore, it is not clear why one would consider human purposes and practices as situated outside the natural domain, unless one has additional metaphysical assumptions attaching to the definition of ‘natural’ such that human activity is by definition not natural and that any kinds which result from such activity cannot be considered natural. Causal accounts such as Boyd’s or Khalidi’s would only accept that this is a real issue if they also accepted the claim that human practices and purposes cannot be responsible for the existence or maintenance of natural kinds. As discussed in the previous chapter both these accounts reject such claims. Slater’s account does not consider a causal story, where it exists, as the determining factor of the natural kind status of the category and therefore this objection does not apply to the account I endorse in this thesis.

3.4 Social kinds as natural kinds

Although kinds such as gender-based categories or emotion are social kinds, nonetheless they are kinds whose members share a property cluster that can be objectively determined. Haslanger provides an analysis of gender categories as socially constructed concepts which primarily track social roles. For example, in the case of the gender kind woman, what all women have in common is ‘being systematically subordinated along some dimension by virtue of being perceived to have a female’s biological role in reproduction’ (Haslanger 2000). According to this view, the kind woman is a natural kind whose members share external relations. There is something that is common between members of the kind woman which can be determined by objective observation of paradigm cases (Haslanger 2012). The idea is that, as all members of the kind occupy the same social role, they will share the same type of relations with society. Although, according to this view, the kind woman is not based on the category of biological sex, Haslanger maintains that it refers to a natural social kind. In other words the fact that such kinds are caused by social factors is irrelevant to whether they will show an aptness for induction and explanation. It is an
empirical matter whether social kinds such as gender categories are important kinds with causal efficacy and explanatory or predictive power (Boyd 2010).

An added complexity relating to some social kinds is that people may have mistaken beliefs as to the causal factors that underlie their existence. For example, in cases of social kinds such as race or gender categories people may mistakenly believe that these categories are biological categories in the same way as tiger, that the clusters associated with them are clusters of biological properties. However, these categories are social kinds; they are not biological categories at all in the same way that tiger or shark are. Revealing the real causes underlying the construction or maintenance of such kinds should not entail that these kinds are automatically non-natural. According to the epistemological view, what determines natural kind status is the aptness for induction and explanation. We may be mistaken as to the reason underlying the stability portrayed by the clustered properties of gender categories but this does not nullify the stability. On the contrary, it reveals the causal process under which these properties are co-instantiated and maintained and, as such, can provide additional explanation for categories that already exhibit a stability.

Revealing the real causes responsible for the existence and maintenance of some social kinds may operate as an external factor causing changes in the properties associated with the cluster. For example, fast-changing societal norms resulted in radical shifts in our concepts of gender categories. The idea that someone can identify as male or female independently of biological sex was a view not many would have been ready to debate a few years ago. However, it is currently a matter of debate (and some controversy) whether gender categories should track biological properties or whether they should instead be roles one identifies with (Jenkins 2016) or roles one is being classified into (Haslanger 2000). According to the latter two views the properties associated with gender categories are covertly constructed by societal norms and practices. For example many people believe that the gender of an individual is decided by biology, whereas in reality biological causes are not underlying this classification. Instead, biological properties are associated with the sex of an individual, which is considered as something separate to the gender of an individual. As the classification of gender is being used to support phenomena such as sexism or misogyny, revealing the underlying causal factors has normative force. It changes the way we perceive these classifications and it provides us with means to address negative social phenomena related to them. This process may lead to a change in some of
the properties associated with the cluster. For example in the case of gender categories, if the classification does not have biological causes, the biological sex of an individual would not necessarily be included in the property cluster. This will bring about a change in the property cluster in the way that was discussed in the previous paragraphs.

Griffiths claims that some covert constructions cannot survive the realization that they are merely our inventions (1997, 146). In the case of gender categories he states that:

Our social practices have been transformed by the growing acceptance that traditional gender characteristics are not the inevitable effects of biological sex (Griffiths 1997, 146).

Griffiths argues that knowledge of the nature of the category by those who use the concept would disrupt the process by which the category is constructed (Griffiths 1997, 147) as individuals classified as ‘woman’ or ‘man’ become aware that these classifications are not based on biological properties. This awareness influences the behaviour of kind members and disrupts the process by which these categories are being maintained. One interpretation of Griffiths’ worry here is that this disruption entails that the categories will not show any inductive success. However this should not necessarily be the case. As Haslanger argues, although gender categories were not originally divisions that existed independently in nature, individuals who are classed as members of gender kinds possess common property clusters because they share the same external relations with society.

Although gender-based categories such as woman or man were constructed by us, the members of these categories share something over and above the fact that they are members of these categories. What is common between the members, is that they are being treated as socially subordinate or privileged qua members of the categories woman or man. According to this view, we can make predictions and generalisations about how members of these categories are treated by society on the basis of membership into gender categories. It is these common external relations which support the claim that gender-

---

44 Here I focus on gender-based categories rather than the natural kind status of the category of gender. Gender-based categories are more analogous to categories of emotion rather than the category of emotion however the same objections are used against the natural kind status of both.
based categories such as *man* or *woman* are not gerrymandered kinds and that their members share commonalities viewed as stable property clusters.\(^{45}\)

It is important to note that Haslanger’s view is not that there are intrinsic commonalities between members of gender kinds. Instead, as Barnes argues, the commonality is to be found in the social structure that explains the phenomenon of gender oppression (2017, 2424). Barnes interprets Haslanger’s view as follows:

> Gender for Haslanger, just is the broader structural feature: the systematic privileging of some and disadvantaging of others based on perceptions of biological sex. That’s what explains what women and men have in common with each other across planes, cultures, times etc. And that’s what explains the way in which societies seem to systematically disadvantage women, even though they do so in strikingly different ways. (2017, 2423)

The case of gender categories shows that some property clusters may consist of or include external relations rather than exclusively intrinsic properties. However, we can still engage in successful induction and explanation on the basis of these relations and as such, such property clusters can be considered natural kinds.

Considerations relating to covert constructions are particularly relevant to the case of *emotion*. *Emotion* does not have an unchanging underlying essence in the same way as the kind *gold*. Interactive kinds such as *emotion*, allow for change in the clustered properties. Its property cluster is subject to the change characteristic of species kinds such as *tiger*. External and internal factors influence which properties cluster with stability. However, if this change happens slowly and with low frequency scientific theory will be able to study it, analyse it and engage in induction and explanation for the property cluster undergoing such change. According to the epistemological view, change within the property cluster is to be expected and it does not preclude that the property cluster will not show any stability. The objection from interactivity therefore is intimately connected to an

\(^{45}\) Philosophers such as Hacking (1995) or Griffiths (1997, 146) call this phenomenon ‘dynamic nominalism’ to differentiate it from simple nominalism which they interpret as the idea that what is shared between members of the category is the fact that they are members of the category. It is not the focus of this thesis to make a distinction between ‘realism’, ‘nominalism’ and ‘anti-realism’. What is important for the argument I am making is whether some type of commonality is shared by the members of gender kinds over and above the fact of category membership.
assumption that the clustered properties are unchanging or that some of the clustered properties are necessary. Under the epistemological view both these essentialist assumptions are rejected.

The claim that covert constructions such as gender categories may be considered natural kinds usually attracts controversy. The reason is that according to the traditional view of natural kinds, they are metaphysically fundamental. Considering a kind to be natural is supposed to attach to it an inevitable and unchanging essence. If gender or race categories are natural kinds then we have to accept them as they are. According to Haslanger the worry seems to be that if we accept that such kinds may be natural kinds, then we are politically constrained to honour and sustain them as they are (Haslanger 2012, 155). However, as I argued in the previous chapter, according to the epistemological view there is no metaphysical fundamentality attaching to natural kinds. It is not necessarily the case that natural kinds have unchanging essences. Instead accepting that kinds such as gender categories or emotion are natural kinds comes with an epistemic requirement that they show an aptness for induction and explanation. Therefore interactive kinds which are also covert constructions can be considered candidates for natural kind status if one does not accept metaphysical assumptions of unchanging essences. In addition, discovering all the external and internal causal forces that are influencing the property cluster is likely to increase the category’s aptness for induction and explanation; however, it should not determine whether the category is a natural kind.

One could argue that in cases where we mistakenly believed that a kind was based in biology but it turns out that it is a social kind, we should eliminate such kinds from scientific talk and investigation. However, as I argued in this chapter, it is not clear what argumentation supports such views. It may be the case that gender categories or emotion are not biological categories; this should not necessarily lead one to support eliminativism of those terms from all scientific vocabularies. Haslanger in the same way claims that not all cases of social construction necessarily lead to eliminativism of the concepts thus constructed. She states the following:

If, however, we allow that we might have been wrong about the kind of thing we were talking about, then we need not take the eliminativist route…..In other words, we would offer a very different construal of the
target concept. Is this shifting the meaning of the term? Or is it discovering the meaning? Often this is exactly the issue at stake between eliminativist and non-eliminativist social constructionists. (Haslanger 2017, 157)

As per the above in cases where the property clusters are stable enough to generate successful induction and explanation, discovering the causal stories associated with them and classifying them in the appropriate scientific disciplines should not entail that the terms we used to refer to them need to be eliminated.

3.5 Summing up

In this chapter I argued that several metaphysical assumptions about the relation between the ‘social’ and the ‘natural’ lead to objections against the claim that normative interactive kinds lacking the same or sufficiently similar causes can be considered natural. According to the epistemological view, some categories are created, maintained or discovered by human minds, decisions, practices, languages or institutions. These practices and institutions are ‘natural’ mechanisms in the sense that they form part of and exist in nature. Mallon in the same sense argues that:

Humans and their minds, cultures, decisions, social practices, languages, institutions, and arrangements are simply natural mechanisms among others that can subserve and sustain real categories.66 (Mallon 2016, 158)

The claim that some social kinds such as woman or man are natural kinds usually causes controversy. This may be due to the fact that some social kinds such as gender categories or emotion are covert constructions. Although these kinds are socially constructed and maintained, they were considered as categories describing biological kinds. Having biological or chemical causes usually comes with the assumption of a status of metaphysical fundamentality in the sense that biological or chemical causes can allegedly be reduced to physical causes and these causes are considered metaphysically

---

66 Mallon focuses on the reality of the category whereas my focus is on whether these categories can be considered as stable property clusters and whether they show an aptness for induction and explanation.
fundamental. However according to recent views, these kinds are not biological or neurobiological kinds (Haslanger 2000; Mallon 2016). Instead they are kinds that involve different scientific domains such as sociology or social psychology and which cannot be reduced to biological or physical kinds.

In this chapter I argued that normative, interactive kinds, membership of which may be caused by multiple factors, can be considered natural kinds. These features may influence the property cluster of some kinds but this influence need not entail that the kind does not show an aptness for induction and explanation. This should be the case even for kinds such as emotion which are considered covert constructions in the sense that the kind was mistakenly believed to be reducible to a neurobiological kind. In the next chapter (chapter 4) I develop an argument for the claim I have been maintaining throughout this chapter, namely that emotion is not a neurobiological kind. Instead, I argue that emotion is an interactive, normative kind in social psychology that has multiple causal factors. Nonetheless, it is a category that shows an aptness for induction and explanation in the way required by the epistemological view.
Chapter 4 – The natural kind status of emotion

In the first part of this thesis I defended the claim that some social kinds can be considered natural kinds. This entails that emotion does not necessarily need to be a neurobiological kind (or a non-social kind) for it to be considered natural. In this chapter I argue that emotion is not a neurobiological kind. In the first part I analyse the main views on the natural kind status of emotion (discussed in chapter 1) to claim that any view which determines the natural kind status of emotion on the basis of the assumption that emotion is a neurobiological kind should be rejected. I argue that identifying emotion as a neurobiological kind ignores important features of emotion which are investigated by disciplines located in the social sphere such as social psychology, sociology or social ecology. I provide evidence for the claim that emotion is a complex process which cannot be reduced to, or analysed in, strictly neurobiological properties.

4.1 Emotion as a single state vs. emotion as a complex process/state

In 1.1 I briefly discussed three different views on the natural kind status of emotion. According to views A and B, all emotions are basic emotions and the kind that includes only basic emotions (and their blends according to view B) as members is a natural kind. These basic emotions are neurobiological kinds; they are identified with clusters of neurobiological properties. According to view C, basic emotions are not natural kinds and, as such, the kind including them is not a natural kind in any scientific discipline. In this chapter I reject all of the above views by arguing that we should not use BET and its postulated kinds (that is basic emotions) to determine the natural kind status of emotion and that emotion is not a neurobiological kind.

Whether emotion is a natural kind depends on how one defines ‘emotion’, on the theory one endorses on what kind of entity emotion is.47 Broadly speaking there are two main different views on the nature of emotion.48 According to the first view, emotion is a

---

47 Here the word entity is used to mean ‘thing’ or ‘phenomenon’ and is not intended to incur any ontological commitment.

48 As Price argues, the division between complex process theories and single component theories is a simplification (2015, 59). However, agreeing with Price, I claim that this division highlights different perspectives from which to approach the debate on the natural kind status of emotion (Price 2015,
complex entity which includes certain fundamental features. Several features are suggested by different theories but most theories consider the following among the features that capture the distinctive character of emotion: a) its phenomenology (the fact that it feels a certain way to have an emotion) b) its intentionality (the fact that emotion is about something) c) its epistemological significance (the fact that we assess emotions from a variety of different perspectives) d) its relation to attention and motivation e) its functional role in dealing with specific types of challenges (Deonna & Teroni 2012, xi; Price 2015, 77). According to such views, what we call emotion includes all or most of these features. Although researchers may disagree on which of these features are fundamental they agree that we are able to individuate emotion from other psychological phenomena, such as memory or cognition on the basis of these features. I call this view ‘the complex view’.

There are several different theories on the nature of emotion that endorse ‘the complex view’ whilst disagreeing on the importance each feature holds for emotion. For example, perceptual theories (see for example Prinz 2004) give more emphasis to the feeling aspect of emotion; cognitive theories (see for example Nussbaum 2004) emphasize the cognitive assessment (judgement) element, whereas hybrid theories (see for example Deonna and Teroni 2012) emphasize more than one fundamental feature of emotion. According to the complex view emotion is a dynamic entity with several different properties such as phenomenological, intentional, functional and evaluative properties.

According to an opposite view, although in everyday language we associate emotion with its phenomenology, its epistemological significance or its intentionality, these features are caused by the emotion, they are not fundamental features of the emotion itself (Adolphs 2017, 27). According to this view, emotion is an internal state of some sort which does not include as fundamental features a phenomenology, an intentionality or some sort of cognitive judgment. According to this view, which I call ‘the simple view’, the emotion state is separate from its phenomenology (the feeling element) or its epistemological assessment, categorisation, evaluation, etc.. As such, emotion should be identified with a state rather than a process; the phenomenology, cognitive assessment (e.g. labelling, 

---

60. As I claim in what follows, single component theories are more compatible with the claim that emotion is a neurobiological kind.

49 There are several differences between the two views but the question that I focus on in this thesis is whether the emotion state includes features of phenomenology, intentionality and epistemological significance as constituents rather than as causal consequences.
categorisation and evaluation of emotion) or behavioural features are derivative from but not constituents of the emotion (Adolphs 2017, 27).

As I mentioned above, in this chapter I argue against all dominant views on the natural kind status of emotion, which are based on the acceptance or rejection of Basic Emotions Theory. In 4.2 I show how this theory has influenced the debate on the natural kind status of emotion and how it is connected to the three different views which I reject. I discuss criticism received against earlier versions of this theory that spoke against predictions made by BET as it was originally formulated. I then discuss attempts to revise this theory so that it accommodates the criticism received. I claim that the revised versions are closer to the ‘simple view’ of emotion, which, as I argue, comes with reductionist assumptions about the nature of emotion.50 The older versions of BET are challenged by empirical data whilst the later versions, by removing several elements from the constituents of emotion, encounter problems with individuating and defining emotion. My discussion focuses on the feature of phenomenology of emotion, that is, the feeling aspect of emotion. I argue that any view committed to the claim that the phenomenology of emotion is something separate to emotion itself (simple view) provides a distorted analysis of what emotion really is. I conclude that we should not use BET to determine the natural kind status of the category of emotion because it presupposes the claim that emotion is a neurobiological kind. Instead I argue that emotion has features that cannot be reduced to neurobiological properties (complex view) and that emotion is not a neurobiological kind.

4.2 BET and the debate on the natural kind status of emotion

Basic Emotion Theory (BET) holds that some emotions such as fear, anger or joy are identified with basic emotions, i.e. automated hardwired responses found universally in humans, controlled by subconscious mechanisms and triggered by specific environmental stimuli (Ekman and Cordaro, 2011).51 According to such views, some of the mental

---

50 It is possible that one can hold the view that the emotion state is something distinct from its phenomenological properties without also endorsing reductionist assumptions about emotion. However as the phenomenological properties are usually considered to be the irreducible features of emotion, if they are separated from the emotion state it becomes easier to argue that the emotion state is reducible to a neurobiological state. In this thesis I focus on views according to which this separation of emotion from its phenomenology is motivated by precisely this possibility of reducing the emotion state to a neurobiological state.

51 Some researchers such as Ekman and Friesen (1986) use the terms basic emotions and ‘affect programs’ interchangeably. In this thesis I use the term basic emotions throughout and argue that
phenomena we consider to be emotions in commonsense psychology such as love or jealousy are not basic emotions. They are different mental phenomena which involve cognitive processes; they are not automated or hardwired. According to one strong version of BET, any mental phenomenon which in commonsense psychology we consider to be an emotion but is not a ‘basic emotion’ does not qualify as an emotion (Ekman & Cordaro 2011, 365).

As mentioned in the introduction, according to Ekman and Cordaro (2011, 365) the following characteristics are found in most basic emotions:

- Distinctive facial signal
- Distinctive patterns of physiological changes
- Automatic appraisal of the stimulus
- Unbidden occurrence; each basic emotion cannot be stopped once activated
- Distinctive subjective experience

These characteristics are parts of the basic emotions. Basic emotions are activated automatically when certain environmental challenges arise.

A common assumption among supporters of BET theory is that the category that includes only basic emotions could be considered a natural kind (Griffiths 1997). This assumption takes basic emotions to be identical to neurobiological phenomena. To give an example, according to this view the emotion of fear is identical to the activation of a specific neural circuit in the brain. Activation of this specific neural circuit is responsible for physiological changes, e.g. heart palpitation, skin conductivity; behavioural output, e.g. running away; and subjective experience, e.g. a distinctive feeling. Basic emotions according to such views are neurobiological phenomena sharing identical neurobiological properties. So in the case of the example used above, all instances of the basic emotion of fear would share the activation of a specific neural circuit in the brain; they would be hardwired and automated.

According to some philosophers, the distinction between basic and non-basic emotions is the biggest threat to the unity of the category (Deonna & Teroni 2012) and, therefore, basic emotions are not emotions. If my argument is sound, any view which takes basic emotions to be identical with ‘affect programs’ will have to concede that ‘affect programs’ are not emotions.

52 This is not an exhaustive list. See chapter 1, section 1.1 for the whole list of characteristics.
is highly relevant to the debate on the natural kind status of the category. Griffiths for example argues that emotion is not a legitimate object of scientific investigation because it includes heterogeneous mental phenomena, i.e. basic emotions and non-basic emotions. He states that “There is no object of scientific knowledge which corresponds to ‘emotion’” (Griffiths 1997, 16). This claim focuses on the epistemic role of the category of emotion and suggests that we will not be able to produce successful generalisations and predictions relating to all members of the emotion category because emotion is not a natural kind.

As discussed in the previous chapter, according to this view, emotion as a category is similar to the category of jade; the category of jade includes two different minerals, jadeite and nephrite. Observations made on instances of jadeite cannot be generalised to instances of nephrite. However, both were included under the category of jade until scientists discovered that they did not share an identical microstructure. Although jadeite and nephrite share similar superficial properties, the category of jade is not a scientific category i.e. it is not a natural chemical kind and, as such, we will not be able to make successful generalisations applying to the members of this category. Griffiths argues that a similar claim of disunity can be made for the category of emotion because it includes heterogeneous phenomena (Griffiths 1997, 14). Examples of such heterogeneous phenomena would be an instance of the emotion of fear and an instance of the emotion of jealousy. The point that Griffiths is making is that there is nothing in virtue of which both instances would be considered members of the same natural kind emotion and that we cannot create successful induction and explanation for a category which includes both.53

4.2.1 BET and defences of the unity of the category of emotion

Several philosophers in the last 20 years have replied to the disunity claim in order to defend the natural kind status of the category. Griffiths’ claim and certain defences against the ‘disunity thesis’ rely on BET and integrate it in their theories. In what follows I discuss some of these attempts and how they relate to the claims made by BET.

a) Basic emotions and their compounds are natural kinds

53 Griffiths is also making a separate claim to the effect that instances of the same folk category of fear, say fear of spiders and fear of an economic crisis, are not members of the same natural kind fear. He claims that we are not able to create successful induction and explanation for the kind that includes them as members. In this thesis I focus on the natural kind status of the category of emotion rather than the natural kind status of the subcategories of emotions.
Although BET was predominantly used to support attacks on the unity of emotion, Scarantino and Griffiths claim that this theory could also provide a way to defend the unity (Scarantino & Griffiths 2011) between some specific members. When they discuss the notion of psychological basicness, which they interpret as the view that a basic emotion does not contain another emotion as a constituent part, they state the following:

As a matter of historical record, most theorists who have defended the notion of psychological basic emotions … have tried to reduce the wide variety of discrete emotions to a more manageable and fundamental subset of discrete emotions. A prospective payoff of this project is that if the component emotions are natural kinds, the compound emotions can also be thought of as natural kinds on account of their components, and the emotion domain acquires theoretical unity underneath the superficial dissimilarities. (Scarantino & Griffiths 2011, 452)

One way one can defend the unity of the category of emotion is to claim that only phenomena qualifying as basic emotions or include basic emotions as components are emotions. If these basic emotions are natural kinds, the composites of basic emotions can also be considered natural kinds and therefore the whole category which includes basic emotions and their composites can be considered natural.

b) Emotion is a neurobiological kind because all emotions share a causal mechanism Charland states that:

My strategy will be to argue that emotion is a neurobiological natural kind for the very same reasons that emotions are natural kinds.54 (Charland 2002, 518)

According to Charland, some individual emotions are natural kinds in what he calls ‘a causal sense’ because they share the same causal mechanism. He claims:

What is central about natural kinds in that sense is the role they play in stating reliable generalisations, and the causal mechanisms in virtue of which they exercise that role. Some emotions can plausibly be argued to be natural kinds in this causal sense. The same is true of ‘emotion’, the kind formed by those basic emotions….Using this causal sense of a natural kind it

54 The context makes it clear that Charland is talking about basic emotions in this passage.
is possible to argue that emotion is a neurobiological kind. (Charland 2002, 533)

The reasoning behind this claim is that basic emotions are neurobiological kinds because they share similar causal mechanisms. Charland seems to think that the same claim can be made for the category that includes these neurobiological kinds as members, emotion. It is not clear what Charland means by the claim that emotion as a category is a neurobiological kind in a causal sense. One way to interpret this is to say that the members of this category share similar causal mechanisms as this is what Charland uses to support the claim that individual emotions are natural kinds. For the purposes of this chapter what is important to note is that Charland integrates BET in his theory of what emotion is and claims that emotion is a neurobiological kind.

c) Basic emotions and non-basic emotions are embodied appraisals

Prinz takes his theory of ‘embodied appraisals’ to be compatible with BET, as basic emotions according to his view are embodied appraisals and non-basic emotions are either blends of basic emotions or they emerge from basic emotions:

There must be emotions in place for these processes to occur. Basic emotions are those materials. (Prinz 2004, 150)

Prinz claims that there is no inelegance in the category that contains both basic and non-basic emotions. He states:

All emotions are states of the same type...All emotions are constituted by embodied appraisals alone... The case for unity is quite overwhelming. When one reflects on the many things that emotions have in common, it is rather surprising that many philosophers have been tempted to divide up the category. It is really no wonder that we have a single English word for the spectrum that runs from animalistic ecstasy to aesthetic delight. Emotions are a

---

55 Another way to interpret Charland’s claim is to say that emotion is a natural kind because its members are natural kinds. This claim is in need of support as it is not clear that a category that includes different types of natural kinds must also be a natural kind. To give an example, depression and schizophrenia may be natural kinds but that does not imply that the category of mental disorders which includes them both must necessarily be a natural kind (See Beebee & Sabbarton-Leary 2010 for more on this point).
natural kind in a strong sense. They share a common essence. (Prinz 2004, 101-102)

As noted in the quote by Scarantino and Griffiths above, the view held by Charland seems to be that emotion as a category is a natural kind because it shares the same causal mechanism in the same way that basic emotions and their blends (non-basic emotions) share the same causal mechanisms. Prinz on the other hand considers BET to be compatible with his own theory as basic and non-basic emotions are, equally, embodied appraisals.

   d) Some non-basic emotions may actually be basic emotions

Deonna and Teroni suggest that the unity of emotion can be preserved because it is possible that the number of basic emotions will increase, if more emotions are identified as basic emotions:

   One cannot exclude the possibility, for example that instances of an emotion dependent on a certain complex judgement share similar underlying causal mechanisms, and therefore qualify as affect programs insofar as they implicate all the physiological and neural changes characteristic of them. (Deonna & Teroni 2012, 25)

The view held by Deonna and Teroni seems to be that if all instances of financial crisis-type fear (Fear 1) share similar underlying causal mechanisms, financial crisis-type fear may qualify as a ‘basic emotion’ (and a natural kind) and therefore, the number of members of the category which includes all basic emotions, so Fear 1, Fear 2 etc., will increase.

These views are examples of two different ways philosophers integrate BET in the debate on the natural kind status of emotion. One can accept the claim that some emotions are ‘basic’ and that these basic emotions are natural kinds. One can then hypothesize that non-basic emotions are blends of basic emotions and as such, that they could also be considered natural kinds because of their constituent parts (e.g. Prinz or Charland). Alternatively, one can predict that the number of basic emotions is going to increase with future empirical research such that some of the emotions which are considered currently as ‘non-basic’ emotions prove to be ‘basic’.56 According to this view the expanded

---

56 Ekman for example accepts this view in his later work where he provides an expanded list of basic emotions (see Ekman & Cordaro 2011).
category of basic emotions could also be considered as a natural kind, because basic emotions are natural kinds.\(^{57}\)

In this chapter I argue that we should not rely on BET to determine the natural kind status of emotion because there are reasons to think that this theory is false.\(^{58}\) Criticism raised against the methodology of empirical studies supporting BET casts doubts on one of this theory’s main claims, namely the claim that facial expressions of some emotions are recognised across cultures (Russell 1994; 2016). Several claims were made assuming this universality feature, such as a one-to-one correspondence between basic emotions and biological markers (e.g. distinctive facial signals or dedicated neural circuits). These claims are challenged by empirical data (Lindquist et al. 2012, 2013; Barrett 2006; Russell 1994, 2016; Touroutoglou et al. 2015; LeDoux 2015, 2016). According to these studies predictions made by BET to the effect that each basic emotion corresponds with the activation of a specific neural circuit and with a package of distinctive physiological and behavioural responses have not been verified. Instead, we have data showing the activation of several neural circuits (LeDoux 2016) and a variety of physiological and behavioural responses (Lindquist et al. 2012; Touroutoglou et al. 2015) corresponding to each basic emotion. Several attempts to revise this theory so that it can withstand the criticism received (see for example Scarantino 2015 or Hutto 2018) reject most commitments made by the original BET and as such, it is not clear that they can be considered a continuation of this theory.\(^{59}\)

In what follows I first discuss one of the main claims made by BET relating to the universality of recognition of facial expressions of emotion and several associated predictions made to explain accumulated data from research conducted in the last 40 years. I then look at empirical data which challenges the predictions made by BET relating to a one-to-one correspondence between each basic emotion and a dedicated neural circuit or patterns of physiological and behavioural responses (Barrett 2006; Russell 1994, 2016; Lindquist et al. 2012; Lindquist et al. 2013).

---

\(^{57}\) An implication associated with this view is that any emotion which does not prove to be a ‘basic emotion’ will not be considered a natural kind.

\(^{58}\) A similar suggestion is made by Colombetti, who uses different arguments against the use of the notion of basic emotions in the debate on the natural kind status of emotion. See Colombetti 2014.

\(^{59}\) Scarantino (2015) attempts to address such worries but the theory he argues for, what he calls ‘new BET’, also faces important challenges, as I argue later in this chapter.
BET generated research on certain features of a specific list of emotional phenomena such as changes in ANS (Autonomic Nervous System) and other physiological changes or on emotional expression. It has contributed greatly in the scientific study of emotion in general and of specific emotions in particular. Recent studies started focusing on what have been previously considered as non-basic emotions and shed further light on whether this distinction can be justified by empirical data or whether it needs to be put to rest.  

In this chapter I claim that many of its predictions have not been verified and the commitments of the revised versions exclude important aspects of emotional phenomena. Psychological theories such as constructionism (Barret & Russell 2015) or componential theories (Scherer 1994) identify emotion as a more complex entity possessing a mix of perceptual, cognitive, and affective properties which is not reducible to any of these whilst retaining its nature and therefore can better accommodate the complex nature of emotion. Agreeing partly with rival psychological theories, I make the following claims: i) emotion is not a neurobiological kind, ii) basic emotions are not emotions, and iii) we should not use BET in the debate about the natural kind status of emotion.

4.2.2 The universality claim and distinct biological markers

According to BET, a small set of specific emotions (basic emotions) is hardwired in our brains. These emotions are programmed responses ready to be activated automatically when certain environmental challenges arise such as dealing with dangers (Fear), removing obstacles (Anger), expelling noxious substances (Disgust), suffering losses (Sadness) and so forth (Scarantino, 2015). These programmed responses have corresponding biological markers in the form of distinctive facial signals or the activation of dedicated neural circuits. According to earlier versions of this theory all basic emotions are characterised by a specific phenomenology, i.e. an inner feeling that sets them apart from other kinds of stimulus-response mechanisms lacking this phenomenology (Ekman & Cordaro 2011). Although there is disagreement amongst supporters of this theory as to which emotions qualify as ‘basic’, the most popular candidates included in the various suggested lists are the ones corresponding to the folk psychological emotion categories of fear, anger, surprise, sadness, happiness and disgust.

---

60 See for example Tracy et al. 2005 who researched the emotion of pride.
As mentioned in the introduction, there are several possible interpretations or senses of the term 'basic'. According to the biological sense, 'basic' emotions are hardwired, that is, they are built in our nervous system. Scarantino argues that:

A hardwired circuit is one that is built in the nervous system, inherited, present at birth, and homologous to the brain circuits present in evolutionarily related species. (Scarantino 2015, 341)

The terms ‘hardwired’, ‘basic’ and ‘innate’ are used interchangeably and point to the biological markers mentioned above. Each emotion is matched with a corresponding neural circuit or, in other words, the activation of the neural circuit is emotion-specific (Scarantino 2015, 363). In what follows I take the term ‘biologically basic’ in relation to emotion to mean an emotion that has evolutionary origin and distinct biological markers in the sense suggested by Ortony and Turner (1990) and accepted by Scarantino and Griffiths (2011, 448). Therefore, according to this theory, some emotions are hardwired in our brain in the sense that we can match these emotions to distinct hardwired biological markers such as distinctive facial signals or the activation of distinct neural circuits.

To get support for the claim that certain response programs are hardwired in humans, studies are performed across cultures to check whether these programs are shared universally. If some emotions are hardwired in human brains, then these emotions should be universal. Scarantino claims that the best evidence we have for the universality of emotions concerns the recognition of facial expression of emotion by perceivers of different cultures (Scarantino 2015, 339). There are several studies that claim to provide support for universal recognition of facial expression of emotion (Tomkins & McCarter 1964; Ekman 1972; Izard 1971; Ekman & Friesen 1986; Matsumoto et al. 2009). These studies, also called ‘judgement studies’, typically ask participants to match a photograph depicting a posed exaggerated facial expression to a list of specific emotions. Other studies (production studies) focus on the production of spontaneous/or posed facial expressions of emotion and their comparison with predicted facial expressions (Friesen 1972; Ekman, Davidson & Friesen 1990; Matsumoto & Willingham 2006; Matsumoto & Willingham 2009).

According to supporters of BET, these studies provide convincing evidence that recognition of facial expressions of a set of emotions is universal. BET attempts to provide an explanation for this universal recognition and suggests that there exist certain hardwired programs in human brains and that facial expressions are parts of the hardwired
programs. BET thus predicts that a distinctive facial expression is part of a discrete hardwired response program. A study by Matsumoto and Willingham (2009) on the production of facial expressions of sadness and happiness by congenitally blind Olympic athletes from different cultures suggested that the facial expressions were similar. This adds support to the claim that people from different cultures produce similar facial expressions when experiencing the emotions of sadness and happiness.

BET theory can be considered as an attempt to provide an explanation for the universality of recognition of facial signals. If therefore it is shown that the universality of facial recognition of emotion is not warranted by empirical data, it is not clear that someone could claim that at least some emotions are universal and biologically based in the sense that they are hardwired through the existence of this specific biological marker. According to Prinz the universality claim is an important premise in what he aptly calls ‘the argument from facial recognition of emotion’ (Prinz 2004, 110). He describes this argument as follows:

Members of different cultures make similar faces under similar circumstances; this suggests they are experiencing the same emotions; if so, the best explanation is that those emotions are biologically based; therefore, some emotions are biologically based. (Prinz 2004, 110)

This argument could be expanded as follows:

1. Members of different cultures produce and recognise similar facial expressions under similar circumstances.
2. The best explanation of 1 is that members of different cultures experience the same emotions.
3. The best explanation of the fact that members of different cultures experience the same emotions is that these emotions are hardwired (biologically based)

So

4. At least some emotions are hardwired (biologically based).

If premise 1 above is shown to be false, the argument is not sound. If emotion were biologically based this would provide support to the claim that emotion is a biological (or neurobiological) kind. I argue that we have evidence that speaks directly against premise 1.
According to BET, facial expressions are viewed as parts of the emotions on a par with other bodily changes that take place when we are experiencing an emotion. Examples of such expressions would be smiling when happy and widening of the eyes when afraid. These are considered to be involuntary bodily changes; for example, Ekman’s study on the Duchenne smile claims that certain types of smile (Duchenne smiles) typically occur when one is experiencing positive emotions, e.g. whilst we are experiencing happiness (Ekman, Davidson & Friesen 1990).

The importance of distinctive universal signals for the theory of basic emotions is fundamental. Ekman originally claimed that there is a one-to-one correspondence between an emotion and a distinctive facial signal; however, he has since updated his view to accept that some basic emotions, such as guilt and shame, are not associated with a distinctive facial signal (see Ekman & Cordaro 2011, 366). Therefore, according to Ekman’s revised version of BET, there is not a correspondence between each basic emotion and a distinctive facial signal. However, where a facial signal exists, its recognition should be universal; it should be universal because this facial signal is hardwired across cultures and, because it is hardwired, people across cultures recognise it. It is this commitment that I discuss in what follows.

### 4.2.3 Challenges to the universality claim

Since Tomkins’ and Ekman’s studies in the 60s and 70s, there has been a sustained critique of both the design methods used in the studies and the interpretation of the results. Critics claim that several design factors affected the predicted outcome of those studies and that the evidence does not show universal recognition of the facial expressions of some set of emotions (see Russell’s criticism in detail in Russell 1994 and 2016; Nelson & Russell 2013). Among the criticisms made against the universality claim are the following:

- The forced choice method used in judgement studies increases recognition rates: The participants were asked to choose from a list of specific emotions instead of choosing from all possible emotions the participants recognise. Subsequent

---

61 One could also explain the co-existence of universal facial signals and universal recognition by claiming that the facial signal served a communicative function. So in this case facial signal and recognition of facial signals co-evolved. I will discuss the communicative features of facial signals in section 4.2.4.
studies show that providing perceivers with more labels or allowing free-label choice decreases agreement (Banse & Scherer 1996; Russell, 1994).

- Free choice tasks do not show a one-to-one correspondence between a facial expression and a specific emotion: After criticism received in relation to the forced choice method used in earlier studies, researchers who endorse BET conducted studies using free choice tasks (Izard 1971; Boucher & Carlson 1980). Those experiments purported to show average to high recognition rates of facial expressions of emotion. However, as the conductors of the studies acknowledged, participants chose very different emotions to match to the facial expressions given. For example they chose different words such as ‘pain’, ‘loneliness’, ‘pity’, ‘worry’ etc. to describe the same facial expression. The experimenters decided to group together these different emotions under a bigger emotion family i.e. ‘distress’ and then treat ‘distress’ as a synonym for ‘sadness’. Prinz draws the following conclusion:

  Free choice experiments simply have not shown that specific emotions, such as sadness, are consistently associated with faces. Rather they show that broad ‘clusters’ of emotions are associated with faces (2004, 112).

Studies conducted using free choice methods show that participants matched single outputs (the same facial expression) with a broad cluster of related emotions. Unless these related but distinct emotions are brought under the same umbrella term, results in these studies speak against BET.

- Exposure to facial expressions of different cultures through movies, music, interaction etc. increases recognition rates: Participants in many studies were college students with significant exposure to facial expressions of different cultures. A meta-analysis found that recognition (matching) scores vary with culture, language and education (Nelson & Russell 2013; Trauffer, Widen & Russell 2013)

- Within-subjects rather than between-subjects experimental design increases recognition rates: The same participants were shown all photos depicting different emotions and then were asked to match these to a specific list. In combination with the forced choice method this means that each participant’s chance of matching a photo with the predicted specific emotion from the list increases after
the first match is made, as the participants know they have to match the next photo to one of the remaining items on the list (Russell 1994, 112).

- Posed exaggerated expressions increase recognition rates: Models were asked to exaggerate the expressions of specific emotions to enhance recognition rates. This means that in the cases where participants choose the predicted outcome, they are not identifying typical facial expressions of emotion but stereotypical facial expressions of emotion. Stereotypical facial expressions are usually found in art or other communicative processes. This speaks in favour of cultural influences and a communicative role rather than proving a biological connection (Russell 1994, 114).

- Statistical averaging errors: In some studies the results were summed up and averaged across all different cultures. This masked low recognition rates in some non-Western cultures because the results were summed and averaged with high recognition rates in Western cultures (Russell 1994, 109).

- Facial expressions are difficult to recognise out of context: Studies show that facial expressions are difficult to match to specific emotions out of context (Jack 2016; Russell 2016). A wide range of information can be considered contextual, such as situational facts (e.g., an athletic contest), the presence of others, the personality and knowledge of the perceiver, cultural display rules, the perceiver’s emotional state, bodily gestures and posture, etc. (Hess & Hareli 2016).

- Arbitrariness of basic emotion lists: Colombetti argues that the original list of nine affects chosen by Tomkins and the list of six emotions subsequently chosen by Ekman in the judgement studies appear to be completely arbitrary (Colombetti 2014). No argument is provided as to why those emotions were chosen and others such as love or jealousy were left out. This arbitrariness has led to a generation of studies on the properties of the chosen emotions. Colombetti rightly points out that we know very little about the physiology, expression and brain processes of jealousy, shame, envy, or love because of the assumption that these are not basic emotions and therefore their manifestations would be variable (Colombetti 2014, 40).

- Recent studies on remote cultures do not provide support for BET: One of the most significant tests for the claim of universality involved conducting studies on non-Western cultures with no prior exposure to Western civilisation (Ekman et al. 1969). However the few studies of remote societies conducted recently showed
a large cultural difference and a weak to non-existent support for BET’s predictions (Nelson & Russell 2013; Russell 1994; Gendron et al. 2014).

Researchers who endorse BET, may claim that on its own each of the issues raised above is not a sufficient reason to cast doubts on the universality of recognition of facial expressions of emotions. However, the combination of these factors provides evidence for doubting the claim that facial expressions are recognised universally as corresponding to a set of specific emotions, namely basic emotions. Russell argues that ‘sufficient evidence has now accumulated to show that BET’s claims about universal recognition from facial expression are unwarranted’ (2016, 165).

4.2.4 Is facial expression necessarily a part of emotion?

BET suggests that facial expressions are parts of response programs activated automatically by specific environmental stimuli; as such, they are necessarily parts of the emotions in the same way that physiological changes (e.g. heart racing, pupil dilation, etc.) or a specific phenomenology are. For example, your heart starts racing when you are afraid, and fear is characterised by a distinctive feeling. As mentioned in section 1.2, a study by Matsumoto and Willingham (2009) on the spontaneous facial expressions of blind athletes at the 2004 Athens Paralympic Games found that congenitally blind judo athletes at these Games produced the same facial configurations of emotion as sighted athletes in similar situations. Hwang and Matsumoto interpret these findings as showing that facial expressions of emotion are biologically innate because it ‘is impossible for congenitally blind individuals to produce by imitation the complicated facial expressions involved in complex muscle combinations activated spontaneously in less than a second when they experience an emotion’ (Hwang & Matsumoto 2016, 144).

However, the critical studies mentioned in section 4.2.3 show that culture and context interfere in the relation between facial expressions and emotion in a way that casts doubt on whether facial expressions are necessarily parts of emotion. One way to explain the mismatch observed between facial expressions and specific emotions is to reject the claim that facial expressions are necessarily parts of the emotion. This relates to a suggestion made by Fridlund, according to which facial expressions are communicative of intended actions and not necessarily expressive of inner feelings (Fridlund 1994). On this view, facial expressions are not necessarily expressions of an inner feeling that we experience but they are a signal to the rest of the world of an action that we are about to perform. An example would be the showing of teeth when angry which, under this view, is a signal
that the bearer may strike (Prinz 2004). This view is related to Frijda’s (1986) suggestion that emotion prepares us for action: it involves motivational states and behavioural tendencies. For example when experiencing fear, the heart rate increases in preparation for flight. Fridlund, in a similar way, suggests that facial expressions are not necessarily expressive of inner feelings; they can be signals of intended action. Importantly, then, the existence of facial expression is not sufficient evidence that someone is experiencing an emotion.

A study exploring the connection between facial expressions and emotions by Fernandez-Dols and Ruiz-Belda (1995) found that Olympic athletes rarely smile just after winning gold medals; they only smile reliably on the podium facing an audience. The results of these studies seem to contradict the results of the studies by Ekman and Tomkins mentioned in the previous section. The issue here seems to be that the facial expression in the studies by Fernandez-Dols and Ruiz-Belda, where the bearer of the smile is likely to be experiencing happiness, seems to be a controlled communicative signal rather than an involuntary bodily reaction. Kraut and Johnston (1979) similarly found that bowlers tend to smile when they turn to face others rather than after hitting a strike.

Prinz interprets these experiments as showing that:

Universal facial expressions may show only that people in different cultures have similar intentions in similar circumstances, leaving the question about emotions unanswered. (Prinz 2004, 111)

One of BET’s predictions was that the facial expression of emotion is simply an automated response, it is a bodily reaction that happens to us; it is unmediated by any complex cognitive process in the same way that other changes in physiology happen automatically when we undergo a certain emotion e.g. increase in heart beat when feeling fear etc. The studies mentioned above by Fernandez-Dols speak directly against such claims as they allow for the possibility that the facial expression is not an automated response, but it is perhaps mediated by more complex cognitive processes such as communicating information to the environment. Relevant to the point made above, several studies show that recognition of facial expression of emotion varies with context (Aviezer et al. 2008; Hassin, Aviezer & Bentin 2013; Carroll & Russell 1996). A facial expression such as smiling can be interpreted as a signal of a specific emotion, e.g. happiness, in one situation and the same facial expression can be interpreted as a signal of a different emotion, such as anger or sadness, in a different situation. However, this
speaks directly against the supposed one-to-one correspondence between a distinctive facial signal and a specific basic emotion.

Ekman attempted to explain this influence by culture and context with the notion of ‘display rules’; these are cultural scripts that either allow or prohibit the display of facial expressions (Ekman & Friesen 1971). Ekman’s account still predicts that facial expressions are necessarily parts of the response program activated. However, they are de-activated by display rules. It is not clear what these rules are – whether they are parts of the emotion or whether they are separate psychological states which interact with the emotion – and I will not assess them in any detail in this chapter (see Russell 1994 and 2016 for further discussion). Arguably the hypothesis of ‘display rules’ goes some way in addressing the variation shown in facial expression across cultures. In the next chapter I argue that even if Ekman is right and ‘display rules’ explain the variation shown in the facial expression of emotion, these can be considered part of emotion rather than an additional element separate to the emotion. For the purposes of this section I note that the notion of ‘display rules’ may provide an explanation for the fact that facial expressions associated with some basic emotions vary depending on context.

What the discussion above makes obvious, however, is that if facial expression is a biological marker for certain basic emotions, this association is different from the predicted association between dedicated neural circuits and certain basic emotions. To see why this is so consider the following; one can experience one of the basic emotions \( x \) without any facial expression \( y \) accompanying the emotion. There are many cases where one can feel profound sadness without exhibiting the facial expression associated with sadness.\(^{62}\) What such cases show is that it is possible to have \( x \) (the basic emotion) without having \( y \) (facial expression ‘marking’ that emotion). On the other hand the association between a ‘basic emotion’ \( x \) and an activation of a specific neural circuit \( z \) is different; we cannot have an instance of \( x \) without an instance of \( z \). Therefore if one wishes to maintain facial expression as a biological marker of emotion, one has to endorse a weaker association between facial expression and emotion. This would mean that in some cases facial expression would be a part of (and a biological marker for) emotion and in other cases it would not be. Someone can experience an emotion yet lack a facial expression,

\(^{62}\) It is fair to assume that Ekman’s theory allows for cases where the facial expression which typically accompanies the emotion of sadness is absent without the need for display rules to explain such absence, e.g. when someone is physically tired or when someone is competing in sport and is so focused that the influence of context (e.g. display rules) seems to disappear.
and someone can have a facial expression which does not express an underlying emotion. To give an example, laughter may arise as an immediate response to being tickled but in this case laughter isn’t necessarily the expression of joy or amusement. At the same time someone can experience an instance of joy without producing a smile (assuming this is the facial expression associated with joy). In other cases one can experience an instance of joy that is accompanied by a facial expression (e.g. smile).

As mentioned above, arguably BET can reply to such worries by using the concept of ‘display rules’ and by loosening the connection between facial expressions and basic emotions. ‘Display rules’ can explain cases where a facial expression would be expected and is activated but is suppressed by other brain processes. Alternatively, BET can reject the claim that facial expression must always accompany an emotion and suggest that this applies only to some instances but not all. The problem with accepting the latter claim is that the concept of biological markers usually refers to a close correlation and not a loose connection between phenomena.

It seems, then, that the claim that the facial expression of emotion is a biological marker of emotion is liable to strong objections. The notion of ‘display rules’ goes some way in addressing the objections related to an observed dependence between recognition of facial expression and context. At the same time the notion of ‘display rules’ reveals an interaction between emotional responses and the social environment. This interaction results in the regulation of emotion (at least in relation to facial expression). As I argue in section 3, this could provide support for the view that emotion is a social kind rather than a neurobiological kind. Finally, the association between a facial expression and some basic emotions seems to be more relaxed than the association between the activation of a dedicated neural circuit and a basic emotion. In what follows I turn to look at other biological markers to see whether they can better support the claim that basic emotions are hardwired.

4.2.5 Basic emotions and dedicated neural circuits

In the previous section I argued that facial expression and more specifically distinctive facial signals do not correspond one-to-one with some specific emotions. It is still a matter of debate in affective science whether empirical data point towards a one-to-one correspondence between basic emotions and biological markers other than facial
expression, such as dedicated neural circuits, automated patterns of ANS activity and physiological changes (Levenson 2011; Lindquist et al. 2012; 2013; Barrett 2006; Thagard & Shröder 2015; Touroutoglou et al. 2012; Touroutoglou et al. 2015). Cacioppo et al. (2000), for example, challenge the prediction that each basic emotion corresponds to specific ANS changes, whilst Kreibig (2010) found that each basic emotion shows a variability of responses. As Robinson suggests, ‘the evidence that there are uniquely identifying bodily profiles for each basic emotion is somewhat ambiguous, especially with regards to autonomic differences’ (Robinson 2018, 67).

Researchers who endorse BET have tried to accommodate some of these findings in at least two different ways. Ekman (Ekman 1999; Ekman & Cordaro 2011) claims that basic emotions should not be identified with folk emotion categories. On Ekman’s revised view, basic emotions do not correspond to the folk categories of emotion. Instead he claims that each of the basic emotions forms a family; for example under ‘anger’ we find the subcategories of irritation, rage and fury. According to Ekman, the following families of emotion are the basic emotions: ‘anger’, ‘fear’, ‘disgust’, ‘sadness’, ‘contempt’, ‘happiness’, and ‘surprise’. Ekman claims that what connects all the related states of a family is the emotional theme. The emotional theme is what is hardwired, what is universal. Irritation or rage under this account are sub-categories of the basic emotion of ‘anger’ which share the hardwired emotional theme. Ekman’s revised account predicts a correspondence between the emotional theme and a dedicated neural circuit with distinct patterns of ANS activity, changes in physiology and behaviour.

According to the above we would expect instances of rage and irritation to share a dedicated neural circuit and similar patterns of ANS activity, changes in physiology and behaviour because these instances share the emotional theme of ‘anger’. However, several studies challenge this prediction, as instances of ‘fear’ and instances of ‘anger’ showed more similarity in neurophysiological activity than two instances of ‘anger’ (Barrett 2006; Lindquist et al. 2013). This is problematic as if all instances of ‘anger’ shared an identical theme, we would expect to find more similarities in neurophysiological activity between two instances of ‘anger’ than between an instance of ‘anger’ and an instance of ‘fear’. Therefore, Ekman’s revision is challenged by empirical data that do not support a one-to-one correspondence between emotional themes and biological markers (Touroutoglou

---

63 As in previous chapters, I use italics and quotation marks to refer to Ekman’s suggested basic emotions to distinguish them from folk emotion categories.

A separate revision is attempted by Scarantino and Griffiths, who agree with Ekman that basic emotions do not correspond to folk emotion categories (Scarantino & Griffiths 2011). However, they also deny that basic emotions correspond to Ekman’s families. According to Scarantino we should not identify basic emotions with the emotion categories we use in commonsense psychology because we have evidence against such correspondence. He states that:

The mistake of traditional basic emotion theorists has been to suggest that the exaggerated expressions associated with basic emotions in the presence of focused/powerful/sudden prototypical elicitors are universally found whenever ‘fear’, or ‘anger’ or ‘disgust’ are activated. This prediction has been widely and rightly debunked by psychological constructionists. (Scarantino 2015, 339)

Scarantino claims that we cannot create the type of generalisations we would need for the categories to be considered natural kinds, if the candidate categories are the folk emotion categories. If we tried to create a generalisation at the level of bodily changes, neural circuits, current function or development, we would be able to find members of any given folk psychological category E to which the generalisation applies and members to which it does not apply (Scarantino 2015, 356). The claim here is that folk categories of emotion do not share enough common properties in bodily changes, neural circuits, current function or development and as such, we cannot create successful generalisations for them.

In addition, Scarantino accepts that the claim made by Ekman, that there is a one-to-one correspondence between emotion families and dedicated neural circuits, lacks empirical support (Scarantino, 2015). He suggests instead that there is a one-to-one correspondence between subcategories of emotions such as BODY BOUNDAry VIOLATION DISGUST and CORE INGESTION DISGUST (related to the emotion Disgust) and a dedicated neural circuit. The difference between Ekman’s revision and Scarantino’s revision is that the suggested concepts of basic emotions pick out different referents. For example Ekman identifies the basic emotion as the emotional theme Disgust whereas

---

64 I use capital letters for Scarantino’s suggested basic emotions to distinguish them from folk emotion categories and from Ekman’s families.
Scarantino identifies basic emotions with the subcategories of Disgust such as BODY BOUNDARY VIOLATION DISGUST or CORE INGESTION DISGUST. Scarantino’s revision (what he calls ‘new BET’) is purported to avoid the challenges raised above relating to output manifestations as it allows for a variability of manifestations associated with the same basic emotion. This hypothesis abandons a requirement for identical output manifestations and accepts that basic emotions can be either output-rigid or output-flexible because they are influenced by the context they arise in. However it maintains the commitment to a correspondence between each basic emotion and a dedicated neural circuit. Scarantino states that:

    Hardwired neural circuits are only expected to be found at a much finer grain of analysis (e.g. what he calls ‘the unconditioned basic fear’ circuit).  
    (Scarantino 2015, 364)

Scarantino suggests that although there are multiple circuits underlying the category of basic Fear, there is one circuit responsible for CONDITIONED BASIC FEAR and a separate one for UNCONDITIONED BASIC FEAR. He states:

    There are multiple circuits underlying basic Fear, so we should distinguish between CONDITIONED BASIC FEAR and UNCONDITIONED BASIC FEAR; there are multiple circuits underlying basic Anger, so we should distinguish between DEFENSIVE BASIC ANGER and OFFENSIVE BASIC ANGER; and so forth. (Scarantino 2015, 338)

According to this view, each revised basic emotion is correlated with a dedicated neural circuit, which is activated just in case one experiences the specific basic emotion. A dedicated neural circuit will be activated every time we experience a specific basic emotion such as BODY BOUNDARY VIOLATION DISGUST or CORE INGESTION DISGUST; this circuit is distinct for each basic emotion (Scarantino 2015, 364).

Scarantino supports his view with evidence from studies in rodents showing the existence of hardwired circuits responsible for orchestrating responses to the sorts of challenges basic emotions were assumed to have evolved to solve (2015, 338). According to

\[\text{65 It is questionable in itself whether 'new BET' can be considered a continuation of traditional BET and so whether Scarantino is justified in calling this theory 'new BET'. This, however, is not the focus of this chapter. Here I raise two challenges for the revised versions of BET and claim that the commitments of this theory are challenged by recent empirical data and lead to the identification of emotion with a state with lacks a phenomenological aspect.}\]
Scarantino, these include distinct circuits for responding to unconditioned threats, responding to conditioned threats, producing defensive aggression, producing predatory aggression, responding to body-boundary violations and responding to repulsive foods (Scarantino 2015, 338). According to this view, each basic emotion could form a neurobiological natural kind whose members share a specific neural circuit. The specific neural circuit will be activated each time we experience this fine-grained emotion. At the same time, under this view instances of the same basic emotion may involve different facial, behavioural, ANS and phenomenal changes:

Second it is no longer assumed that the output of basic emotions must necessarily be a rigid cascade of mandatory responses….In the general case, each basic emotion will be associated with prioritised response tendencies geared toward solving a specific fundamental life task in a context dependent way. (Scarantino 2015, 364)

Scarantino states the following as the commitments of ‘new BET’ (Scarantino 2015, 363):

1. Basic emotions are evolutionary adaptations, selected for because they are efficient solutions to fundamental life tasks.
2. Basic emotions are associated with programs considered as specific patterns of responses.
3. Basic emotions are associated with hardwired neural circuits, but such circuits do not correspond one-to-one with any folk psychological emotion category.
4. Basic emotion programs are elicited by automatic appraisals and be either output – rigid or output flexible. They can be output flexible in the sense that facial, behavioural, autonomic and phenomenal changes can be manifested in different ways depending on context (Scarantino 2015, 364).
5. Basic emotions are pancultural, presented across species and emerge early in development.
6. Basic emotions are designated by theoretically motivated subcategories, such as UCONDITIONED BASIC FEAR, CONDITIONED BASIC FEAR, BODY-BOUNDARY VIOLATION DISGUST, INGESTIVE BASIC DISGUST, DEFENSIVE BASIC ANGER and so forth.
As mentioned above, these emotions are considered ‘basic’ because they are biologically basic in the sense that they are innate, or hardwired. It is not clear from the above what Scarantino means by ‘associated with’ in point 3. One way to interpret this is to understand this association as an identity relation. Scarantino states that this association is not a one-to-one correspondence with folk psychological categories, but what is important for the purposes of this chapter is that he predicts that the association depicts a one-to-one correspondence between a distinct neural circuit and a finer grained emotion. As quoted above, he claims that such a one-to-one correspondence is still predicted by new BET by stating:

Hardwired neural circuits are only expected to be found at a much finer grain of analysis (e.g. the unconditioned basic fear circuit). (Scarantino 2015, 364)

The quote above allows the interpretation that Scarantino’s new BET predicts a one-to-one correspondence between the activation of a neural circuit and a ‘finer grained’ basic emotion.

From the revised list of commitments provided by Scarantino, it appears that the only commitment to a biological marker relates to point 3, that is, each basic emotion has a one-to-one correspondence with the activation of a distinct neural circuit. This is the only remaining commitment associated with biological markers, which Scarantino hypothesizes we will be able to observe in lab experiments in the future. Here the claim is that there is (and can be observed) a correspondence between the activation of a distinct neural circuit (say Neural Circuit 1 or NC1) and a type of basic emotion (say E1).

In what follows I argue that we have sufficient evidence to doubt that this prediction will be verified. Instead, we have empirical data that speak directly against it. As this is the last commitment associated with biological markers, I argue that we do not have robust evidence to support the claim that basic emotions are biologically basic in the sense that they are innate or hardwired because they share the activation of a distinct neural circuit. In addition, I claim that ‘new BET’, by reducing basic emotions to restricted, finely grained psychological phenomena, needs to provide a good argument to support the claim.

---

66 This is different to the trivial claim that a neural circuit is activated every time one undergoes emotional or mental phenomena in general. Here it is the one-to-one correspondence between a specific category of emotion e.g. UNCONDITIONED FEAR and a specific neural circuit e.g. NC1 which is the important claim.
that these phenomena are emotions. Unless one presupposes that emotion is a neurobiological kind and that its property cluster includes exclusively neurobiological properties, one should reject the claim that basic emotions (both the old and the revised ones) are emotions.

4.2.6 Challenges to the claim for one-to-one correspondence between basic emotions and dedicated neural circuits

Attempts to revise BET such that it identifies emotions with affective states that are finer-grained than the folk psychological concepts of emotion attract two different challenges. Firstly, the insistence of ‘new BET’ that each basic emotion corresponds with a dedicated neural circuit is problematic. Evidence from neuroscience shows that each emotion is correlated with the activation of several neural circuits which are not exclusively associated with emotional phenomena (LeDoux 2015; 2016). Secondly, identifying emotions with affective states of narrower content runs the risk of distorting the nature of emotional phenomena by leaving out important features of emotion e.g. the phenomenological aspect; this creates problems with individuating emotional phenomena.

a) On the one-to-one correspondence between each basic emotion and a dedicated neural circuit

LeDoux claims that we have evidence that some of the defensive responses associated with emotional responses, such as freezing, are hardwired (LeDoux 2015; Panksepp 1998; 2012). However, he also states that:

Feelings of fear and anxiety arise from circuits in the brain that differ from the circuits that control the expression of defensive behaviours like freezing, and are likely vulnerable to different factors…Certainly the circuits that control defence responses and give rise to feelings of fear do interact, but this does not mean that they are the same. (LeDoux 2016, vii)

In the quote above LeDoux claims that different circuits are responsible for responses to the ones that are responsible for the feeling of fear when people experience instances of CONDITIONED FEAR. There is not one dedicated neural circuit which is correlated with the emotion of CONDITIONED FEAR or UNCONDITIONED FEAR but more than one neural circuit for each basic emotion. As LeDoux argues, the feeling aspect of the emotion is correlated with the activation of a separate neural circuit to the one
correlated with, for example, the defensive response of freezing. This speaks directly against ‘new BET’s’ requirement that each basic emotion corresponds one-to-one with a dedicated neural circuit. This challenge applies to Ekman’s emotional theme of Fear and also to Scarantino’s finer suggested basic emotion of CONDITIONED BASIC FEAR, since they both hold a commitment of a one-to-one correspondence between a dedicated neural circuit and a basic emotion. There is activation of more than one neural circuit when we are experiencing instances of CONDITIONED BASIC FEAR, for example. This fits better with the observation that emotion appears to be output-flexible in terms of physiological changes or behaviour. As the criticism studies against BET show, each basic emotion corresponds with a variety of different autonomic, facial, behavioural or phenomenal changes (Lindquist et al. 2013). The activation of multiple circuits is likely to correspond to multiple outputs.

If Scarantino were to revise his view so that more than one neural circuit corresponds to each basic emotion, perhaps this objection would not apply. However, this would mean that new BET would have to let go of the commitment to one-to-one correspondence between a refined basic emotion and a neural circuit. Another way to avoid this issue would be to retain the one-to-one correspondence but claim that what we refer to by ‘neural circuit’ is actually a set of individual neural circuits. In such a case each basic emotion e.g. UNCONDITIONED BASIC FEAR would correspond with NC1 (feeling aspect circuit) and NC2 (defensive circuit). This may be possible, but according to new BET, these neural circuits are dedicated to basic emotions. I understand this to mean that they are not activated unless one experiences an emotion. However, if LeDoux is right, the defensive circuits can be activated in the absence of an emotion. This means that at least one of the circuits usually associated with emotion can be activated in the absence of emotion, as happens in the case of instinctual reactions. As such, it is not clear that new BET can retain a one-to-one correspondence even between a basic emotion and a set of dedicated neural circuits.

In the first part of this chapter I focused on facial expressions of emotion and the claim that they form part of the response program associated with basic emotions. The evidence provided by LeDoux indicates that the responses we regularly associate with emotions could be parts of emotion but do not entirely constitute those emotions. This is precisely because, according to LeDoux, there is not one dedicated neural circuit for each specific emotion. Facial expression can be part of a response circuit associated sometimes with emotion and sometimes with other mental phenomena. Studies supporting such views
have shown that facial expression need not be exclusively tied to emotion; for example some types of smiling are connected with interacting with others (Crivelli, Carrera & Fernandez-Dols 2015; Fernandez-Dols, Carrera & Crivelli 2011; Fernandez-Dols & Ruiz-Belda 1995; Fernandez-Dols, Carrera & Barchard 2003). This provides evidence that responses that are sometimes associated with emotions, are correlated with separate neural circuits and function sometimes as communicative means rather than as an expression of inner feelings.

In the case of fear, the emotion is correlated with the activation of more than one neural circuit, e.g. what LeDoux calls the ‘defensive circuit’ and a separate one. LeDoux concludes the following:

I therefore agree with basic emotion theorists that some responses that occur when people feel emotions are hardwired, but I do not think that the emotion, the feeling is hardwired. It is imputed, assembled, constructed, or otherwise cognitively created. (LeDoux 2015, 460)

In the passage quoted above LeDoux seems to hold that the emotion process includes feeling. This can be accommodated by the ‘complex view’, which takes emotion to be a process rather than a single state lacking a phenomenological aspect.

If one insists, as Scarantino and Griffiths seem to do (Scarantino & Griffiths 2011), on a one-to-one correspondence between a basic emotion and a dedicated neural circuit, there is a choice to be made as to which neural circuit corresponds to the basic emotion: the one correlated with the response (in the case of fear, the defensive response), the one correlated with the feeling, or a different neural circuit entirely. To use fear as an example, BET is only compatible with one of the options mentioned above; thus it restricts the content of fear to the defensive response or the phenomenological aspect (feeling) of fear or to something else that is separate from both of those aspects. As ‘new BET’ is committed to the correspondence between each basic emotion and a dedicated circuit, it cannot accommodate a disjunction of these circuits. This is because a disjunction would still be committed to an identity relation between the basic emotion and a single neural circuit. If LeDoux is right in claiming that the same defence circuits are activated when we undergo different emotions such as fear and anger, this would mean that according to the disjunctive account, in cases where the neural circuit activated was the defence circuit without a phenomenological aspect, we would not be able to individuate one emotion from another, e.g. fear from anger, because the defence circuit is associated with both.
To see why this is so consider the postulated basic emotion of UNCONDITIONED BASIC FEAR. If it correlates only with the activation of the survival circuit, it seems that one can have a ‘fearful’ response without feeling fear. However, we can think of cases where one has a fearful response, e.g. freezing, without actually experiencing the emotion of fear, for example when someone is startled. This creates issues with classification. In the absence of the phenomenological aspect of emotion, it seems we cannot distinguish whether the activation of the defensive circuit e.g. freezing is associated with fear or startle. This provides support to the claim that the commitment to one-to-one correspondence creates issues with the individuation of emotion. As such, we can think of cases where we would not be able to determine which emotion an individual experienced if we only look at identical neurobiological properties. In the example used above, the activation of the defensive response circuit may be associated with the emotion of fear or the emotion of surprise or startle. In such cases we would need to know the feeling associated with this phenomenon to identify the emotion as fear or startle or surprise.

b) Identifying emotion with narrow states that are different to folk categories of emotion and lack phenomenological or behavioural features

Scarantino’s view predicts that we will be able to create one-to-one correspondences between a basic emotion (revised concept) and a dedicated neural circuit; that is, on the basis of empirical studies we will be able to create generalisations of the following type: Whenever neural circuit \(x\) is activated, the participants experience basic emotion \(y\), e.g. UNCONDITIONED BASIC FEAR. In the empirical studies mentioned in the previous section, such generalisations are usually based on observations of i) the neurophysiological activity of the participants (e.g. via fMRI, EEG), ii) the reported experience of the participants, and iii) the behaviour of the participants whilst they undergo various tasks (e.g. Izard’s experiments). In the studies on facial expression mentioned in the previous section, the participants are asked to report the emotions they recognise or experience on the basis of folk categories of emotions such as fear, anger, disgust, etc. (e.g. the studies conducted by Ekman or Izard). The concepts the participants use to perform these tasks are used as the basis for the correspondences made between the neurophysiological activity observed and the emotion reported by the participants.

An objection that can be raised here to Scarantino’s view is that, if the suggested subcategories of fear are associated with different neural circuits and do not map onto the folk psychological category of fear, it is not clear on what basis one can claim that they
are subcategories of fear and not of some other category without involving circularity. To establish correspondences between distinctive neurophysiological responses (NC1, NC2,…) and distinctive emotions (E1, E2, …) we need an independent way of distinguishing E1 from E2. However, once emotions (‘basic’ or ‘non-basic’) are dissociated from the folk-psychological categories, one way E1 and E2 can be distinguished is through the difference observed between NC1 and NC2. In such cases E1 and E2 simply collapse into NC1 and NC2. However, unless one endorses a reductionist account of emotion, one cannot simply identify NC1 with E1. Furthermore, if the activation of NC1 is associated with more than one emotion – e.g. the freezing response (activation of defence circuit) can be associated with fear or startle, – it seems that the correspondences we decide to make between NC1 and fear only appear arbitrary.

In addition, and as argued by LeDoux, there is activation of multiple neural circuits during an emotional episode. This means that if we insist on a one-to-one correspondence between the activation of a distinct neural circuit and a basic emotion, we have to choose which neural circuit – NC1, NC2, etc. – to match to the basic emotion E1. However such a decision does not accurately reflect what goes on when one is experiencing an emotion. Here, because new BET predicts that the new basic emotions can be output-flexible – e.g. in the ANS changes or output behaviour – depending on the context they are experienced in (Scarantino 2015, 364), we cannot use other evidence of physiological changes to establish this one-to-one correspondence because new BET allows for variation in outputs.

To conclude: Neuroscientific evidence does not support the claim that fear corresponds one-to-one to a dedicated neural circuit. Instead, fear is correlated with the activation of several neural circuits. If one wishes to identify the emotion of fear (the finer grained type of UNCONDITIONED BASIC FEAR) with the activation of one of those circuits, as new BET seems to hold, then we would need an additional argument to support the choice of one of the activated circuits over the others. If one chooses to identify emotion with the neural circuit associated with the freezing response, this would create issues with the individuation of emotions. This is because the defensive response of freezing is not a necessary or sufficient feature for the emotion of fear or the emotion of startle. It is therefore not clear on what basis Scarantino’s revised concept of UNCONDITIONED BASIC FEAR rather than CONDITIONED BASIC FEAR or CONDITIONED ANGER, etc., corresponds to the activation of NC1, because all these could correlate with the activation of NC1.
In addition, new BET, by narrowing the content of basic emotions, implies a conceptual disparity between what participants report during the experiments, which are based on folk emotion concepts, and the scientific interpretation of what emotion is being experienced. It implies that we should not use folk categories of emotion as scientific categories. As such, there will be a conceptual disparity between what participants report or recognise during the experiments, which is based on folk-emotion concepts, and the scientific interpretation of what emotion is being experienced. Unless there is a conceptual revision on a very large scale, this disparity will always affect hypotheses that predict one-to-one correspondences between neurophysiological activity and specific emotions whilst at the same time rejecting folk categories of emotion. Furthermore this revision does not accurately reflect the activation of neural circuits in emotional experience unless one presupposes that we should identify NC1 with E1, i.e. unless one presupposes a reductionist view of emotion.

The claim that some emotions correlate with the activation of dedicated neural circuits that do not include the feeling aspect of emotion is compatible with the ‘simple view’ according to which emotion should be identified with a state that does not include several features we associate with emotion in commonsense psychology (e.g. phenomenology). Therefore, although BET, as originally formulated, identified some emotions with patterns including a phenomenology, new BET is compatible with views that claim that the phenomenological aspect is separate to the emotion; it does not constitute part of it.

The arguments given above on the nature of specific emotions have implications for the category of emotional phenomena. According to new BET, basic emotions should be identified with states of narrow content which may be separate to features such as the phenomenology, physiological changes or behaviour typically associated with emotions. These narrow states can be identified with the activation of dedicated neural circuits (the one-to-one correspondence claim). I argued against such views on the basis that they distort the nature of emotional phenomena without providing sufficient reasons for doing so. In addition, I argued that, because they strip out fundamental features of emotion, they cannot provide good explanations of emotional phenomena. Rejecting these claims allows for the possibility that all emotions are complex states or processes which should not be separated from features typically associated with them e.g. phenomenological aspect. As these claims are used as a basis for the distinction between basic and non-basic emotions, their rejection provides support for the unity of the category of emotion.
4.3 What is an emotion?

In the previous section I argued that one of the most influential psychological theories on the nature of emotion, BET, has received considerable criticism, and that we have evidence against predictions made by this theory. In light of the criticism received, we should not use BET – either as originally formulated or in its revised form (new BET) – to determine the natural kind status of emotion. In this section I discuss two alternative views on the nature of emotion: a) the view that emotion should be identified with a single neurobiological state and that emotion can be reduced to neurobiological properties, and b) the view that emotion should be identified with a complex process which includes neurobiological, physiological, psychological and non-psychological properties.

In the introduction to this chapter I mentioned two different views on the nature of emotion. According to the ‘complex view’, emotion is a complex process/event with several different components including phenomenological, appraisal and cognitive features etc. According to this view emotion is a dynamic entity, a process that includes a number of different states that may be causally linked together (Robinson 2018, 51). As Robinson claims, emotion appears to be complex and dynamic:

> Emotions are not instantaneous but take time, and different components in the emotion process are typically on different timelines. In my view, when philosophers claim that emotions are appraisals or attitudes or whatever, they are simply focusing on what they take to be the essential feature of an emotion. But this is a mistake. Emotions are essentially processes involving a number of components. (Robinson 2018, 52)

Robinson claims that emotions are processes and that some of these processes include phenomenological or cognitive features as components (2018, 63). The older versions of BET are compatible with the ‘complex view’ in that they consider such aspects to be parts of emotion. However, the latest versions seem to be restricting the features of emotion by associating it exclusively with the activation of a dedicated single neural circuit. As such, they are compatible with what I mentioned in the introduction as the ‘simple view’.

According to what I called the ‘simple view’, emotion is a single state which can be identified with neurobiological properties; it does not include any non-neurobiological features. In what follows I first discuss the simple view and discuss some findings that
support this view. I then turn to discuss the complex view to argue that it provides better explanations of emotional phenomena. More importantly for the claim I am making in this thesis, it allows for the possibility that emotion is not a neurobiological kind but a social kind: that is, emotion includes social properties that cannot be reduced to neurobiological properties.

4.3.1 Emotion should not be identified with a neurobiological state lacking a phenomenology

A recent proposal by Adolphs (2017) suggests that affective neuroscience identifies emotion with a neural state that is separate from the experience or the conceptual processing of emotion. Adolphs takes the experiential and conceptual aspects of emotion to be caused by, rather than constitutive of, the ‘emotion state’. According to this view, we need to dissociate the emotion state from emotional experience, labelling or concepts (Adolphs 2017, 29). It is not clear what Adolphs means by ‘emotion state’ but what is important for the purposes of this chapter is that he considers this separate from both the experience and the behavioural aspect of emotion. He claims that determining the nature of emotion ‘is an empirical result and not something that we can conflate in our research program at the start’ (Adolphs 2017, 30). He states:

To summarise how people get confused about what is meant by ‘emotion’: there are distinctions between the functional emotion state (the emotion state), its conscious experience (the experience of the emotion)….our ability to think and talk about emotion (conceptualizing emotion) and the behaviours caused by an emotion state (the expression of emotions, emotional reactions). I think emotions are first and foremost about the first of these, and all the others are derivative. (2017, 27)

Adolphs claims that the field of neuroscience should ‘use a scientific concept of ‘emotion states’ that is not based on conscious experience’ and that is separated from the experience or the conceptual processing of an emotion (Adolphs 2017, 27). Adolphs implies that we need to come up with a completely different scientific vocabulary which is different from the one we use in everyday language and the one we have been using in affective science

---

67 Adolph suggests that the following aspects of processing would be central to the functional emotion state: a) emotion-cognition interactions b) emotional learning and memory and c) eliciting strong emotions with ecological stimuli (2017, 28)
so far. This is similar to Scarantino’s suggestion that basic emotions do not correspond to folk categories of emotions.

According to such views, dissociating emotion states from emotional experience, labelling or concepts will help us achieve progress in the scientific study of emotion. One of the reasons motivating such views is the difficulty in isolating the neural correlates of conscious experience (Adolphs 2017, 29). It is difficult to discover identities between neural activation and conscious experience. If we combine this difficulty with the claim that we can have unconscious emotion, one can further claim that the conscious experience of emotion is not a part of emotion. According to this view, emotion is primarily unconscious; it becomes conscious when it is experienced. However, the experience of the emotion and the emotion itself are two different phenomena. Adolphs, therefore, suggests that we identify emotion with a state, which is dissociated from the feeling or behavioural aspect of an emotion.

This suggestion appears to be compatible with new BET’s commitment to a one-to-one correspondence between each finer-grained basic emotion and a dedicated neural circuit, if one interprets such correspondence as implying identity. As we saw above, LeDoux claims that the emotion of Fear is correlated with the activation of several neural circuits. Adolphs suggests that we should not identify the emotion state with all the activated neural circuits, including for example the one correlating with the feeling aspect of emotion, because he considers all but one of the neural circuits associated with emotion to be effects of the emotion rather than constitutive of the emotion.

Damasio’s research may be used to provide support for a dissociation between being in an emotional state and the awareness of being in an emotional state. He claims that we can have unconscious emotion; we can be in an emotional state without feeling anything (1994, 156). He also allows that we can feel as if we are in an emotional state in the absence of bodily changes associated with emotion (1994, 156). These findings support the view that the feeling aspect of emotion can be dissociated from bodily changes and that we can be in an emotional state without experiencing any feeling (1994, 156). Dissociation cases provide support for such views by showing that in some cases some of the features we

---

68 This view may stand in direct opposition to LeDoux, who claims that we can identify the feeling of emotion with the activation of a neural circuit. This opposition shows that the issue of convention raised by Carver in chapter 2 when identifying neural mechanisms may also apply to identifying neural circuits.
associate with emotion seem to be absent; this proves, according to such views, that they are not necessary parts of emotion.

Adolphs supports his view with cases of patients showing a dissociation between the emotion state and emotion concepts (2017). In those cases it can be shown that having conceptual knowledge about specific emotions cannot induce emotional states. This is a description of patient S.M.:

Patient S.M. has bilateral lesions to her amygdala; she shows dissociation with respect to fear that is about as good as dissociations get. She can laugh, she can cry and she can endorse feelings of happiness and sadness and most other emotions. But she does not show any of the effects of a state of fear that we would normally use in order to attribute fear. She does not show normal avoidant behaviours to threatening situations, she does not show autonomic responses or give subjective ratings of fear to normally fear-inducing stimuli, and she fails to show learning based on unconditioned fear in Pavlovian fear conditioning. A subset of the same deficits (minus the subjective ratings and with simpler behaviours) is seen in animals with bilateral amygdala lesions (Adolphs 2017, 27-28)

Adolphs claims that patient S.M. has a concept of fear and she can talk appropriately about fear in conversation but that she does not instantiate the state of fear. He concludes that this shows that the emotion state is separate to conceptual knowledge one has about emotion states.

Even if one accepts that conceptual knowledge about the emotion can be dissociated from the emotion, however, it is more difficult to accept that the same applies to the phenomenological or motivational features of emotion. As I argue in the following section, if one wishes to claim that a particular neural state should be identified as ‘the emotion state’ and that this emotion state lacks a phenomenological aspect, one needs to provide an argument for the claim that the best analysis of emotion excludes phenomenal, intentional, conceptual and behavioural properties from counting as parts of the emotion. I claim that the ‘complex view’ provides a better analysis of emotion and it provides support for the claim that emotion should not be identified as neurobiological state.
4.3.2 Emotion as a complex process/state including phenomenological or behavioural features

There are at least two different ways to claim that emotion should be identified as a complex phenomenon which includes phenomenological or behavioural aspects. According to the first, emotion is a process made up from different states and events. According to the second one, emotion is a state constituted by various events and processes. Goldie argues that the view that emotion is a complex process with several constituent parts is better equipped to provide an analysis of some emotions such as grief. According to this view accounts that privilege one component (be it a state or an event) of the emotion over all others and identify it as the emotion do not provide an accurate analysis of emotional phenomena. Goldie uses the example of grief to argue that emotion is the whole process made up from states and events, which unfolds in a characteristic pattern (Goldie 2012, 61). Goldie states:

Some mental phenomena are primarily processes, and only secondarily can we properly comprehend the mental states and events they are made up of: the parts do not even come into view as parts unless and until they are seen as parts of a particular kind of process. The process is thus ontologically and epistemically prior to the parts. (2012, 61)

As per the above, emotion is a process whose parts are states and events.

On the other hand Soteriou claims that it is possible that emotions are conscious states ‘that are constituted, at least in part, by various events and/or processes’ (Soteriou 2018, 80). Soteriou claims that emotion is likely to be a state that is constituted by events and/or processes rather than a process. In what follows I argue that, although Soteriou and Goldie’s views are different, they both seem to point to the fact that it is not easy to reduce emotion to a simple phenomenon such as a neurobiological state that is separate to other events and processes e.g. its phenomenology or behaviour usually associated with it.

One of the reasons motivating the ‘complex view’ comes from accepting a functionalist account of emotion states. One could interpret a functional state as individuated by its causal profile viewed as the instantiation of a particular pattern of inputs (what causes the state) and outputs (what the state causes). To give an example used by Smith:
For example, the inputs definitive of joy are likely to include positive evaluation of some object or event, the outputs are likely to include the exhibition of a genuine, or Duchenne, smile. This is not to say that every instance of joy causes smiling. Rather, it is to say that joy is partly defined by its tendency to cause smiling. (Smith 2018, 146)

So in the example of joy, let’s say that the process is the following: I suddenly see my brother after 5 months, I positively evaluate this event (input): I feel joyful, my heart races, I smile (output).

Lewis argued that although we need to specify mental states functionally by looking at inputs and outputs, we should identify mental states with the state that realises the functional role even in cases where the usual inputs and outputs are missing (Lewis 1980). Lewis provides a functional account of the mental state of pain as an example of a functional state where we can still claim that a subject is in the functional state of pain even when the usual inputs or outputs associated with this state are missing. This is because the state ‘pain’ is the state that typically realises the functional role.

However, if we identify a mental state as the ‘realiser state’ – the state that typically (but not necessarily always) plays the relevant functional role – then we seem to lose what seems to be central features of many mental states, especially emotions, namely their phenomenological and behavioural aspects. For example, consider the case of Mary who is in a state of pain (she is in the relevant realiser state) but who feels no pain and shows no aversive behaviour, or the case of Michael who is jealous of his brother – he is in the state that typically realises the functional role of jealousy – but neither feels jealousy towards his brother nor exhibits any jealous behaviour. In the first case excluding the feeling of pain seems to remove the state of pain altogether. It seems very difficult to justify the claim that the realiser state in this example is a state that does not include a phenomenological or behavioural aspect without assuming that there is some underlying pathology. In respect of the second example in the absence of subjective experience and behavioural output, it is not clear on what grounds one would identify the realiser state as a state of jealousy rather than some other emotion. As mentioned in section 1 of this chapter attempts to individuate emotional phenomena without the help of the phenomenological and behavioural aspects run the risk of circularity. It seems that unless one accepts a reductionist account where the emotion is identified and reduced to the
single neurobiological state that is separate to phenomenological or behavioural features, one encounters problems with the identification and individuation of mental states.

Consider cases where someone judges the bear threatening without feeling fear towards the bear and without exhibiting fearful behaviour. It seems strange to claim that the person actually is in a state of fear because they judge the bear to be threatening, if they do not experience a state of fear which prompts them to act one way or another. Conversely we can think of cases where one feels fear towards something that is not evaluated as being threatening, e.g. being afraid of oranges. Here, although the person does not evaluate the orange as threatening, we are still likely to describe this state as a state of fear (although perhaps unjustified). These cases suggest that we find it difficult to describe a state as an emotional state in the absence of the phenomenological and behavioural aspects; on the other hand, we can think of many cases where the absence of the relevant evaluation of an object is not considered as a reason not to describe the state as emotional when the phenomenological or behavioural aspect is present. What this shows is that the connection between the several cognitive, phenomenological, physiological and behavioural features of emotion is not easily distinguishable.

Soteriou (2018) in a similar way claims that we should not attempt to functionally characterise some mental states without making any reference to the phenomenal events/processes in virtue of which those functionally specified states obtain. He states:

One might attempt a functional characterisation of the kind of psychological states that obtain when a subject is in pain without making any reference to the phenomenal events/processes in virtue of which those functionally specified states obtain. But in the case of the kind of mental state that obtains when one undergoes an excruciating sensation of pain (i.e. the state of one’s being in pain), if one doesn’t make any reference to the kind of phenomenal event/process that occurs, then, in a crucial respect, one will have underspecified the nature of that state…. And what it is like for a subject to be in pain, on any given occasion, will be determined by phenomenal properties of the phenomenal occurrence she undergoes when she feels pain…. We might say, then, that there is an interdependency between state and phenomenal occurrence when a subject is in pain. (2017, 83)

Soteriou uses the mental state of pain as an example of a phenomenal mental state. He suggests that the same would apply to other phenomenal states such as emotion, and he
claims that there is an interplay between the state and the phenomenal and bodily occurrences associated with conscious emotions. According to this view, there is not always a dividing line between constituents and mere effects of an emotional state (Soteriou 2018, 89). Soteriou’s insight could provide support to the view that emotion should be identified with a complex process rather than a single state and that this process is influenced and interacts with the phenomenological aspect of emotional experience.

Soteriou discusses cases of ‘pain asymbolia’, a condition where patients report that they feel pain without being in pain (Soteriou 2018, 80). Such patients will usually feel a bodily sensation located somewhere in their body, can report on the features of this sensation, recognise the sensation as pain and successfully track objective physiological conditions of their body. However, such an experience does not induce any avoidance, affective or protective reaction. It is difficult in such cases to claim that pain functions as a signal of threat or damage to the organism as the patients do not change their behaviour in any way. Soteriou claims that in such cases pain no longer serves its primary biological function (2018, 81). However another way to interpret such cases is to claim that these patients are not in pain so there is no function that the mental state they are in is supposed to serve. This may be possible if patients miscategorise the sensation as pain when in fact it is some other mental phenomenon.

It is not clear that the cases of dissociation mentioned above support the claim that emotion is a single neurobiological state separate from phenomenological, cognitive or behavioural features. If emotion is an evolved adaptation which provides us with successful strategies in specific situations, we need to analyse all of its components so as to provide better explanations of emotional phenomena. For example, the motivational aspect associated with emotion seems to be intimately connected (in a way which some people may characterise as a cause and effect relation) with the feeling aspect of emotion; it is because fear and anger feel certain ways that our bodies are readied for action and we are motivated to act in one way or another. In the absence of the right type of feeling the relevant bodily states are not activated. What this shows is not what Adolphs claims that cases of dissociation show. Unfelt or uncategorised emotions do not fulfil their biological function because they do not motivate/lead to the appropriate outputs. However, such cases are not the typical cases of emotional phenomena; they are not the rule. They are errors in the process that do not allow the emotion to function properly so that it is successful. Soteriou’s pain asymbolia shows exactly that. The phenomenology of pain is involved in motivating the response of the organism to bodily damage. When the
phenomenology aspect is absent, it is likely that the organism will not be motivated to stop or avoid bodily damage. Cases of dissociation are cases of malfunction, not cases of normal function.

If someone wishes to claim that emotion is a dynamic entity with several occurring components whilst at the same time retaining the assumption that emotion is a functional state of some sort, Soteriou’s view seems unavoidable. However, as Smith argues, supporters of the view that emotion is a state rather than a process or an event comprising several events or processes need to explain how a static entity (e.g. the emotion state) could have occurring parts whilst being static (2018, 147). Although this debate is not the focus of this thesis, it is important because it reveals two different attitudes on determining the nature of emotion. On the one hand, according to the ‘simple view’ emotion is a state that is distinct from its phenomenology, and perhaps also from other cognitive features we associate with emotion in our everyday lives; this state is by definition static in that no parts of it can occur in time. On the other hand, on the ‘complex view’ emotion is either a state (Soteriou’s view) or an event (Robinson’s view) but it is not separate from its phenomenological and cognitive features.

In this chapter I argued that we should prefer the complex view because it provides better explanations of emotional phenomena and helps with the individuation of emotional states. In addition, excluding the phenomenological or behavioural aspect leads to issues with links to motivational states and the relevant appropriate action emotion is evolved to bring about.

4.4 What emotion is not

BET makes several predictions relating to biological markers such as distinctive facial signals and dedicated neural circuits that have been challenged by empirical data (Lindquist et al. 2012; 2013; Barrett 2006; Russell 1994; 2016). A revision by Scarantino (2015), i.e. ‘new BET’, which aims to address those challenges, is committed to a one-to-one correspondence between each ‘basic emotion’ and a dedicated neural circuit. I argued that this contradicts evidence from neuroscience which points towards the activation of

---

69 Soteriou does not use the terminology of parts-whole but rather talks about grounding of the emotional state. He states that ‘we can keep a component view of emotions resisting attempts to pare down the complex constellation of mental and bodily phenomena associated with any given emotion until we identify a single state that is the emotion’ (Soteriou 2018, 89). Here I claim that his view is much closer to my account of emotion as a complex entity rather than a simple entity lacking a phenomenological aspect.
multiple circuits when people experience emotions (LeDoux 2015, 2016; Lindquist et al. 2012; Touroutoglou et al. 2015; Thagard & Shröder 2015). The one-to-one correspondence commitment runs the danger of identifying emotions with psychological states lacking a phenomenological aspect; this creates additional issues with the classification process, i.e. the process of distinguishing emotions from one another and from other psychological phenomena. In addition, I claimed that since ‘new BET’ suggests a dissociation between basic emotions and folk categories of emotions, it is in need of an additional argument which explains the arbitrary matching of the activation of distinct neural circuits to some distinct emotions. Since ‘new BET’ predicts that the revised basic emotions are output flexible (e.g. ANS changes, behaviour), it is not clear what drives the pairing of a neural circuit with a basic emotion unless the latter is identified as the former.

BET was very influential in the debate on the natural kind status of emotion as philosophers of emotion integrated it in their theories to make claims for or against the natural kind status of the category. We have data that speaks both for and against the older versions of BET; we have evidence which shows that a number of emotions come in relatively recurrent patterns of neural and ANS activity (Colombetti 2014, 36). However, we do not have evidence that binds these patterns to hardwired affect programs. The new versions of BET restrict the content of emotion to fine-grained states or clusters of neurobiological properties which are more similar to instincts than to emotions. They are removed from the folk categories of emotion in a way that creates difficulties for their description, definition and individuation. They lead to the claim that emotion is a state separate from its phenomenology and, as such, they do not provide an accurate analysis of emotion. If BET is indeed the biggest threat to the unity of emotion, its rejection calls for a new investigation of the natural kinds in affective science. Colombetti similarly argues that the notion of basic emotions is ‘redundant and misleading, and affective scientists would be better off if they dropped it’ (Colombetti 2014, 40).

Other current psychological theories such as psychological constructionism or componential theories provide better accounts of the nature of emotion by allowing for an alternative option which suggests that several circuits correspond with the emotion

---

70 By rejecting the one-to-one correspondence between basic emotions and folk categories of emotion.
and that these circuits are not dedicated to individual emotions.\textsuperscript{71} Such views see hardwired responses as parts of emotions along with other psychological ingredients such as language, memory, ANS changes, and appraisal processes etc. (Russell & Barrett 2015; Lindquist et al. 2012; LeDoux 2015). Componential and constructivist theories maintain a connection with commonsense concepts of emotions and could be able to provide better explanations of emotional phenomena.

In the next chapter I suggest an account of emotion as a complex entity which cannot be reduced to its constituent parts whilst retaining its nature. I argue that emotion is a stable property cluster or SPC, in the sense explained in Part A of this thesis, and can therefore be considered a natural kind. This property cluster is not a neurobiological kind in that its properties are not strictly neurobiological properties. According to this account, the phenomenological aspect of emotion is part of the emotion property cluster. This cluster includes social properties, which cannot be reduced to neurobiological or neurophysiological properties. This view accommodates empirical data that show variance in physiological patterns associated with emotions because, according to this view, emotion is multiply realised at the physical level. Its property cluster includes properties which cannot be reduced to neurobiological properties such as the activation of distinct neural circuits. At the same time this view allows for some degree of similarity in the physical properties within the emotion property cluster where we focus on specific groups realising emotional phenomena, e.g. by conducting studies among similar cultures or similar species.

\textsuperscript{71} Such views allow cross-cutting between emotions and circuits so that Fear for example may be correlated with activation of circuits 1 and 2 whereas Anger may be correlated with activation of circuits 1 and 3.
Chapter 5 – Emotion as a social natural kind

In this chapter I argue that emotion is a social natural kind. As mentioned in chapter 3, according to the epistemological view all natural kinds are socially constructed because they depend on the practices, methods and classification purposes of individual theoretical domains. However, some of those socially constructed natural kinds refer to social phenomena such as gender categories or emotion. I put forward the claim that emotion is a social kind in that some of the properties included in its stable property cluster are best explained and investigated by several domains within the sphere of the social such as sociology, social ecology and social psychology. As I argue in section 2, we are able to provide successful explanations of emotional phenomena when we consider emotion to be a social kind.

5.1 Emotion, induction and explanation

According to my account, although the emotion property cluster may include some neurobiological properties, these are neither necessary nor sufficient for the instantiation of emotion. It is a contingent fact that the property cluster of emotion includes neurobiological properties such as the activation of neural circuits because emotion is multiply realised. This entails that emotional phenomena can be considered members of the same natural kind even if they have different physical realisations depending on the system which realises them. To give some examples, in human animals the majority of neuron cells are located in the brain whereas in invertebrates e.g. in octopuses they are spread throughout the body (Godfrey-Smith 2017, 105). In addition we can think of possible cases where emotion is realised in the absence of any neurobiological properties, e.g. realised by physical systems that are completely different to our own such as artificial systems. As such different organisms or entities could realise emotion states by instantiating different physical properties. This is compatible with studies conducted on invertebrates such as bees, which exhibit anxious (fearful) behaviour although they have very different brains to human brains (Bateson et al. 2011).

At the same time emotional phenomena can be multiply realised even within the same entity. In other words different systems or neural circuits in the brain at the neurobiological level could give rise to instances of anger such that that emotion state is
realised by different neurobiological properties within the same organism (Lindquist et al. 2013). If emotion is multiply realised, such variety in the physical properties that realise it entails variation between instances of emotion and the physical properties, e.g. neurobiological properties that correlate with it. As mentioned in chapter 4, empirical studies provide support for the claim that we cannot create successful induction and explanation between distinct neurobiological properties and distinct emotions.\(^{72}\)

A consequence of the claim that emotion is a social kind is that the regularities that generate successful induction and explanation will be observed at the functional level and not at the physical level.\(^{73}\) The functional level of explanation of psychological phenomena focuses on the function performed by psychological phenomena (Kim 2011). In the first section of this chapter I discuss the function or task of emotion, which is to enable us to navigate our environment in ways which achieve our goals. According to my view, emotion is a channel through which we interact with the environment and through which we appoint significance to some objects and phenomena rather than others. Following Millikan I use ‘function’ in an etiological way to mean purpose (Millikan 2002). According to this view the function of something is what that thing is supposed to do, which in turn amounts to what earlier things of its type have done that contributed to their survival and reproduction, the doing of which explains the current presence of the thing (Kingsbury 2006, 23).

In what follows I first provide support for my claim that emotion is a social kind by discussing some important social properties of emotion such as the ability to interact with the social environment. I argue that emotion facilitates social life and that some of its properties change as a result of interacting with the environment whilst others only develop because of such interaction.\(^{74}\) The first part concludes with a discussion of the apparent passivity of emotion, the claim that emotion happens to us involuntarily in a way usually associated with hardwired internal processes, e.g. the automatic pumping of blood

\(^{72}\) In chapter 4 I argued that emotion should not be identified as a neurobiological kind. Although emotion has features that can be best explained by analysing neurobiological properties such as the chemical properties of neurons, it cannot be reduced to these chemical properties.

\(^{73}\) Therefore, if we cannot create successful explanation and induction for emotion at the physical level, this does not imply that emotion is not a natural kind.

\(^{74}\) Here by social life I mean something akin to Aristotle’s idea of publicness or Heidegger’s idea of being-in-the-world. I will provide more detail about this idea in section 1 (see also Aristotle 2007, On Rhetoric, 1378a; Heidegger 1953, 134).
by the heart. I show why emotion appears to be automatic and how this is compatible with the claim that emotion is a social kind. In the second part of this chapter I argue that the emotion property cluster facilitates successful induction and explanation. I give examples of induction and explanation by focusing on the social properties within the cluster. In section 3 I give a summary of the claims made in this thesis and conclude that if these claims are true then it follows that emotion is a social natural kind. I finish by making some brief comments on the implications of my conclusion for the study of affective phenomena.

5.2 Emotion as a social kind

In this section I discuss several important features of emotion that support the claim that emotion is a social kind. These include: i) the function of emotion, ii) the irreducibility of emotion to specific properties within the cluster, iii) the multiple realisability of emotion, and iv) the importance of the social environment for the development of emotion. In section 5.2.1 I focus on the interaction between emotion and the social environment to claim that the social environment plays a very important role in both the generation and the regulation of emotion and that these two processes need not necessarily be considered as separate processes. In section 5.2.2 I assess the claim that emotion is an involuntary phenomenon which happens to us out of our control and discuss how this claim can be used against the view that emotion is a social kind.

i) The function of emotion
One way to define the function of emotion is to claim that emotion is an adaptation used to appoint significance, in the sense of appointing value, to objects in the environment. For the organism to be able to appoint significance to stimuli in the environment, the organism must be equipped with a system, process or mechanism that enables it to appoint such significance. Emotion performs this function: it enables the organism to appoint significance to stimuli received both from the environment and the body so that the organism is motivated to act one way or another depending on the organism’s priorities. For example, when survival of the organism is a priority, the organism will develop strategies so that it is motivated to act in such a way that survival rates are maximised; for example, when one feels fear one will run away from danger, e.g. one may see a tiger shark approaching and swim away or get
on the boat. These strategies become available to the organism through learning and reinforcement from the environment.

The emotion property cluster includes several properties that reliably co-exist without any of them being necessary or sufficient. To say that emotion is a property cluster with no necessary or sufficient properties for the instantiation of emotion entails that emotion cannot be reduced to any property included in the cluster. In this way my account is a non-reductive account in that emotion cannot be reduced to any phenomenological, intentional, physiological or any other type of property within the cluster. In a similar way in Tooby and Cosmides’ evolutionary account, emotion is a combination of what they call ‘effects’; it cannot be reduced to any individual ‘effect’ (Tooby & Cosmides 2008). They state:

An emotion is not reducible to any one category of effects, such as effects on physiology, behavioural inclinations, cognitive appraisals, or feeling states, because it involves evolved instructions for all of them together, as well as other mechanisms distributed throughout the human mental and physical architecture. (Tooby & Cosmides 2008, 118)

In addition the cluster itself is not a necessary and sufficient condition for instantiating emotion. As the importance falls on the function performed by the cluster rather than the cluster itself, this means that the cluster as a whole is not a necessary and sufficient condition for instantiating emotion. In other words something could exhibit all properties in the cluster and not be an emotion and conversely something could exhibit some of them and be an instantiation of emotion.

ii) the multiple realisability and the irreducibility of emotion

According to Sznycer et al. (2017) emotion allows the organism to respond to environmental problems by producing evolved adaptations. The way these adaptations are physically realised can differ because each evolved system, e.g. each individual

---

75 Price (2015) in a similar way argues that the function of something is something that ‘it is supposed to do to contribute to something else’ (43). For an interesting discussion of advantages and disadvantages attaching to the claim that emotional responses have functions see Price (2015, 45-55).
emotion, is produced in accordance with the problem the organism chooses to resolve (Sznyyer, Tooby & Cosmides 2017, 56). Sznyyer et al. state:

The systems that give rise to anger, shame, pride, gratitude, fear, jealousy, love, lust (etc.) will not have a uniform architecture defined by necessary and sufficient features. The architecture of each evolved system should reflect the computational requirements of the adaptive problem that selected for its design. (2017, 56)

I interpret the passage quoted above as allowing that the systems realising emotions may be designed in different ways depending on the challenges the organism faces. This view agrees with what I have been arguing in this thesis in the sense that it allows for multiple realisation of emotional phenomena. However, the differences we observe between instances of emotion at the neurobiological level, such as the differences between an instance of fear and an instance of joy, should not entail that the category of emotion that includes them as members is not a natural kind. This is because regularities are not expected to be found at the level of physical realisation if instances of emotion are multiply realised; instead, regularities are expected to be found in the task/function being performed by each emotion. In a similar way to evolutionary accounts, my account allows that individual emotions are different strategies available to the organism when it encounters significant stimuli in the environment.

iii) The development of emotion

The account of emotion I endorse has similarities with Aristotle’s account of emotion and its importance for a successful public life. Aristotle argued that emotion is necessary for living a social life (what he calls ‘publicness’) (2007, On Rhetoric, 1378a, p.113). One needs to learn how to ‘use’ one’s capacity for emotion so that one flourishes, one lives a fulfilled life. According to Aristotle emotion depends on learning and needs to be developed; it is imperative that such learning and development starts from a very early age (1998,

---

76 The same would apply to individual categories of emotion such as fear, pride, guilt etc. One could claim that although instances of individual emotions such as two instances of fear may not share necessary and sufficient properties at the implementational level, they share the performance of the same task, which is to enable the organism to appoint significance e.g. the value of being threatening to objects in the environment in accordance with her several goals and priorities.
Nicomachean Ethics, 1103a, p.29). As our capacity for emotion needs to be developed, the success of our interactions with the environment depends on the priorities of the organism on the one hand, and the cues it receives from the environment during the development period on the other. To give an example, a person needs to learn to feel angry at the right time, in the right way, to the right degree in the right circumstances towards the right person (1998, Nicomachean Ethics, 1106b, p.38). According to Aristotle’s view, this process is learned or could be learned.77

The valuable insight provided by Aristotle relates to the importance of emotion for social life. It is through emotion that we are able to interact in social environments; at the same time these environments are important for the development of emotion. Each individual needs to undergo learning from a very young age so that she develops her ability for emotion in ways which will provide advantages in the social environment she is embedded in. If Aristotle is right in claiming that emotion needs to be developed, this development is essential so that it performs its function successfully. Importantly it needs to be developed within a social environment; it cannot develop outside such an environment. Another way to say this is to claim that emotion needs to be constructed in a social environment.78

5.2.1 Emotion generation and the social environment

I claimed above that emotion develops through an interaction with the environment and that it is open to change and reinforcement from the environment. This allows for flexibility, openness and change, and it presupposes interaction. The significance we appoint to each stimulus by using emotion changes as our environments and priorities change.79 In what follows I argue that the environment is important not only for providing...

77 For more detail see Aristotle’s idea of finding the ‘mean’ in emotional performance which includes both feeling and action (1998, Nicomachean Ethics, 1103b-1105b, pp. 30-35).
78 Different social environments are likely to develop different emotions. This is compatible with the variation in emotion concepts, terms and experiences observed in isolated or radically different societies in the studies mentioned in chapter 4.
79 Most human and many non-human animals, e.g. chimpanzees, gorillas, bees and ants, live in social environments and as such by environment I primarily refer to the social environment. As soon as one accepts that emotion is an evolving system that is shaped and developed by the environment in tandem with the individual, the claim that emotion cannot be defined purely in neurobiological terms becomes more convincing.
stimuli or regulating the emotion but also for the generation of emotion. It is precisely because emotion involves a continuous entanglement between the individual and the social environment that emotion is a social phenomenon. The social environment plays an important role in emotion generation through an interaction with the individual such that i) an analysis of emotion must include an analysis of its social aspects, and ii) emotion generation and regulation need not necessarily be considered as separate processes if emotion is a social kind.

One form of interaction between ourselves and the environment is through the stimuli we perceive. However, this association between a stimulus and a set of responses is open to change through learning in early development (Greenwood 2015). For example, some people are afraid of or disgusted by snakes or tarantula spiders and some are not. We have many examples of individuals who keep snakes as pets without feeling fear or disgust towards them. In addition we have examples of indigenous tribes in South America, e.g. the Piaroa tribe who consider tarantula spiders a delicacy and hunt them from a very young age (Bonta 2017). Such cases provide evidence for the claim that the association between stimuli and emotional reactions is formed by and adapts to the environment one lives in. At the same time these associations can change even after they are formed. For example, one can be raised to be afraid of dogs because one’s parents were afraid of dogs for whatever reason; however as soon as one has the chance to spend some time with a dog the association between dogs and fear may change. Spending time with a dog may get one to form the belief that a dog is not likely to attack or bite someone unless it is provoked. This provides support to the claims that i) these associations are not hardwired, and ii) they are cognitively penetrable: they can be affected by someone’s beliefs after they are formed (e.g. as soon as one forms the belief that a dog is only dangerous if provoked, one may not consider the dog as threatening at all times).

An alternative form of interaction between ourselves and the environment is the fact that the environment provides or reinforces models of emotional strategies when we encounter certain stimuli. For example, depending on which society one lives in, there are different norms and practices associated with the emotion of grief, attaching to when it is appropriate for someone to feel grief, how intense this feeling should be, how long it should last, what behaviours it should be associated with, and so on. A third form of interaction, which is important for my claim, is by determining the meaning of the
situations we are in and guiding us to interpret feelings arising in these situations as
different types of emotion. Therefore, the environment influences emotional phenomena
in several ways, including emotion generation and regulation.

As discussed in chapter 4, one persisting feature of studies conducted on emotion
expression and emotional experience is the variability shown at the implementational
(physical) level of analysis (Lindquist et al. 2013). In addition there is variability in emotion
categorisation and in the concepts used to refer to emotional phenomena (Russell 1994).
This variability is evidence for emotion being formed in tandem between the individual
and the environment. As so, several social properties, such as interactivity, learning and
social development, which are associated with the emotion property cluster, need to be
included in an investigation of the nature of emotion (Greenwood 2015). It is by
analysing these types of properties that we can see why emotion shows an aptness for
induction and explanation, as I discuss in the second part of this chapter. In other words
emotion cannot be analysed in the absence of the social environment that develops it.
This capacity for interactivity, flexibility and change in accordance with environmental
factors highlights the dynamic structure of emotion. Colombetti in a similar way claims
that emotion is the capacity of the organism to form complex dynamical patterns that has
been shaped by evolutionary and developmental time (Colombetti 2014, 82). These
dynamical patterns are flexible and loosely assembled, can be culture specific but may also
recur reliably across cultures.

As mentioned in chapter 4, we have empirical findings against the claim that the facial
expression of some emotions are universally recognised and produced (Lindquist et al.
2013). Some of these findings showed a variability in emotion expression which goes
against the claim that the recognition or production of the facial expression of some
emotions is universal. In addition, variability was shown in the physiological changes

80 Greenwood in a similar way argues that emotion operates primarily in social economies and that it
includes continuously changing neural, corporeal and sociocultural resources (Greenwood 2015,
205).

81 Although I agree with Greenwood’s statement that emotion should be construed as a unified
natural class the function of which is to enable social life (2015, 212) I disagree with her claim that
what she calls ‘precursor emotions’ that is inborn affect mechanisms are basic emotions. Rather,
according to my account whatever inborn mechanisms combine to create emotions are not
emotions themselves as I argued in chapter 4.

82 A fair assumption to make is that the more similar or interrelated cultures are the more similarities
they will show in emotional experience and behaviour.
correlated with instances of individual emotions such as instances of fear or anger. Ekman’s version of Basic Emotion Theory addressed the findings showing a variability in emotion expression by postulating ‘display rules’, rules which society provides and the individual uses to regulate emotion and more specifically the expression of emotion (1972b). He argues that these cultural rules influence emotion expression and, as such, Basic Emotion Theory could accommodate the variation in expression observed in some studies. He maintains that cultural rules do not influence emotion generation.

Ekman therefore accepts the claim that some of the phenomena we associate with emotion depend on culture rather than a distinctive human biology (Ekman 1972b, 212). However, these are restricted to the elicitors of emotions, the behaviour arising from emotion, and the rules of facial expression, e.g. the rules which dictate what facial expression an individual will exhibit in different environments (Ekman 1972b). According to Ekman’s ‘neuro-cultural’ theory the emotion is first generated or triggered, and then the display rules regulate the type of facial expression that will be exhibited by the individual. Ekman explicitly states that this process is sequential that is first the emotion is generated and then it is regulated by display rules. Here I reject this view for two reasons: i) it considers emotion generation to be a different process to emotion regulation, but we have evidence which speaks against the view that emotion generation and regulation are two separate and sequential processes; and ii) the social environment influences emotion generation not only by providing the stimuli but also by providing an interpretation of situations and feelings as well as providing continuous feedback during the interaction with the individual.

One phenomenon which provides support to the claim that the generation of emotion is influenced by the social environment is the fact that some emotions are unique to certain social environments, e.g. the Japanese emotion of amae or the emotion of amok observed in South-East Asian societies (Averill 1986, 98). Amae relates to the emotion of dependency shown by adults akin to what infants feel toward their mothers (Prinz 2004, 131). It is difficult to find evidence of people experiencing this emotion in other societies. Amok is an aggressive frenzy involving indiscriminate killing while in a trance-like state...

As I mentioned in chapter 4, several researchers criticise the idea of display rules because it is not clear what these are and whether they are uniquely associated with emotion (Barrett 2006b; Russell 1994). Here I focus on the claim that such rules influence emotion regulation but not emotion generation.
until the person experiencing *amok* is killed (Averill 1986, 97). The experience of *amok* was associated with a ritualised response to certain socially defined situations (Averill 1986). Interestingly Europeans living in South East Asian societies have experienced *amok* but we do not have evidence for South East Asians living in European countries experiencing *amok*. One possible explanation of this emotional phenomenon is that this emotion is dependent on the social environment one lives in. Unless one lives in this specific social environment one may not experience amok.⁸⁴

Allowing that emotion is open to environmental feedback and change makes it easier to explain the variability of individual emotions between different cultures and the variability shown in the lower levels of psychological explanation such as the implementational (physical) level. Variation in the physiological patterns or emotional behaviour is to be expected in different environments if the cluster properties within the emotion property cluster are produced and maintained by, and adapt to, different social environments.

In addition, according to relational or social interactionist accounts of emotion in social psychology, emotion generation and regulation can be considered a single process rather than two consecutive ones, that is, emotion is not generated first and then the individual regulates it (Campos et al. 2004; Burkitt 2018). Instead there is a constant dialogue between the individual and the social and cultural context one is embedded in, which not only results in the regulation of an emotion already generated, but in influencing what (and if) emotion will be generated and in what way. Herschbach and Bechtel (2015) provide further support for this view and argue that the emotional process is not necessarily a linear process, that is, emotion generation does not happen before emotion regulation in a sequential manner. They suggest that ‘living organisms with homeostatic tendencies operate in non-linear fashion; by the use of negative feedback, they are able to control a mechanism so as to maintain a target state’ (2015, 37).

---

⁸⁴ Researchers who endorse new BET such as Scarantino or Griffiths could accommodate such evidence as they do not claim that all emotions are basic emotions and are thus universal. However Ekman’s BET version could not accommodate such findings as according to his theory all emotions are basic emotions and therefore universal. As such, according to Ekman’s view *amok* would not qualify as an emotion. I discuss the difference between these views in chapter 4. Here I focus on the support that such local emotions provide for the claim that the social environment plays a fundamental role in emotion generation.
The view mentioned above shares a very important principle with the account of emotion I endorse in this thesis, namely that emotion is a complex and dynamic entity or process. Emotion involves interaction between the individual and the environment without this interaction necessitating a linear process where emotion generation is separated from and is followed by emotion regulation in a sequential manner (Herschbach & Bechtel 2015, 41).

Studies in social psychology cast doubts on the claim that emotion generation and emotion regulation are two separate processes (Varga & Krueger 2013). Varga and Krueger explored the relation between emotion generation and emotion regulation in relation to depressive states and child development. They claim that in close relationships such as the ones formed between the child and the caregiver it becomes very difficult to separate the process of emotion generation from emotion regulation and that emotion is generated by both the child and the caregiver (Varga & Krueger 2013, 287). Burkitt supports this claim and suggests that in the close relationship between an infant and a caregiver, the infant ‘does not only learn from the caregiver but rather experiences the world through them, assimilating their bodily stance and the vitality of the bodily interactions between them, through which the infant comes to relate to the world from their own position’ (Burkitt 2018, 172). According to such views, young infants rely heavily on the intervention of caregivers to regulate their emotions through physical interactions such as directing a child’s attention, touching and holding them, using mutual gestures and vocalisations. However, during this process, the feedback the infant receives from the environment does not only concern the regulation of emotion but also the identification of emotion, and its binding to appropriate narratives, interpretation and behaviour. This process takes place from infancy throughout the development period; the child learns to identify, regulate and use its emotions through an ongoing interaction with the environment (Varga & Krueger 2013). This learning period is important for the individual’s interaction with the environment. As Burkitt states, ‘this creates a feeling of connectedness to the world that if lacking can lead to a range of problems in adulthood, such as depression and other negative background emotions’ (Burkitt 2018, 172). As such, the environment influences the identification, interpretation and regulation of emotion.

In addition, in the absence of feedback from the social environment the infant not only lacks an ability to regulate emotion; it also lacks the ability to generate emotion in
situations where this is deemed appropriate. The claim here is not that infants do not have feelings but that these feelings will only be constructed as emotions whilst interacting with the environment. Having only the phenomenological feature associated with the feeling aspect of an emotion does not suffice for having the emotion. An implication therefore associated with social interactionist accounts is that in the absence of the influence and guidance of the social environment the individual will not be able to interpret any feelings or generate and regulate emotions so as to successfully navigate social environments.

To accept that emotion is constructed in tandem between the individual and the environment is to accept that social and political considerations need to be addressed so that we correctly analyse and explain the emotion. Such analysis or explanation of emotion cannot be provided simply by looking at patterns of physiological changes. We need to interpret these patterns by looking at the social context into which they arise and the meanings available to the individual who is experiencing them (Scheman 1980, 32-33). Scheman’s insight concerning the importance of the social environment for the generation of emotion is relevant to the issue under discussion here. According to Scheman, environmental forces influence emotion generation in addition to emotion regulation.

Scheman shows how this process of influencing emotion generation takes place by analysing the role of consciousness-raising groups as a result of the feminist movement (1980). She claims that consciousness-raising groups enabled women who were living in oppressive environments to identify the feelings that arose in them in certain situations as feelings of anger rather than, or alongside, feelings of guilt or shame. Such groups have been criticised as being manipulative of the feelings of women in them because people come to feel what the group deems proper or fitting in a situation, e.g. by bringing about and reinforcing the beliefs that i) in this situation one is being oppressed and ii) when one is being oppressed one feels anger. Scheman accepts that these groups influence the feelings of women in them (1980, 31); the social context created by consciousness-raising groups influences and guides the feelings of the members so that they experience anger when being oppressed in situations where in the past they would have experienced shame or guilt. Scheman’s view is that the interpretation of the feelings as anger or shame importantly depends on the interpretation of the situation as one of oppression.
Therefore the interpretation of the situation as one of oppression influences whether and what emotion will be generated.

In addition, the social environment determines what emotions are deemed appropriate or fitting to that situation and therefore what emotions the individual should feel when oppressed. As Scheman argues, consciousness-raising groups reveal what happens generally in society; social contexts do more than simply determine the situations individuals find themselves in; they also determine i) how these situations will be interpreted, ii) whether such situations should give rise to feelings, and iii) how any feelings arising in the individuals themselves in these situations will be interpreted. The difference is that consciousness-raising groups do it honestly or openly whereas in society most of this process happens covertly.

If the above claims are true and emotion is a social kind, unless an individual interacts e.g. via learning, practice and reinforcement with the environment during the development period, she will not be able to generate or regulate emotion. Although she may experience feelings associated with bodily phenomenology these feelings on their own cannot be identified with emotions. These considerations provide support to the claim that the environment influences emotion generation not only by providing the stimuli and the display rules as Ekman claims. Instead they provide support to the claim that emotion generation happens in tandem between the social environment and the individual.

5.2.2 Emotion generation and passivity

Psychological theories of emotion such as BET are compatible with the claim that the environment plays some role in emotion, however this role is limited to providing stimuli or influencing the regulation of emotion (Ekman 1972a). According to such views, the emotional process is linear; emotion generation happens before emotion regulation in a sequential manner. Emotion generation is perceived as a passive phenomenon, as something that happens to us. If emotion is an automatic bodily response produced in specific circumstances, any social properties included in the property cluster seem to be additional extras, which are associated with the regulation of emotion rather than the generation of emotion. This automaticity provides support to the claim that emotion is a
hardwired reaction in the same way as sneezing for example, and as such, it can be analysed purely in neurophysiological or neurobiological terms. In what follows I provide a reply to this view to show how an account of emotion as a social kind can accommodate its apparent passivity.

The view of emotion as a social kind is in tension with the widely shared intuition that some emotions appear to be automatic or passive, to happen to us out of our control. For example when Britney is insulted by her boss, she is furious and punches the door in anger whilst her boss is still in the room, unable to control the emotion of anger which happens to her involuntarily. Britney cannot control the generation of the emotion of anger because this emotion was activated automatically (e.g. through the activation of a dedicated neural circuit) when she perceived her boss insulting her. Emotions are automatic in a similar way with other bodily involuntary movements, such as sneezing, because they are hardwired (e.g. we come to life with emotions already ‘installed’ and ready to use). According to this latter view although Britney could control how she expressed the emotion of anger (regulation of emotion) she could not control the fact that she felt angry (the generation of emotion).

According to this view the emotion of anger should be identified with the activation of this automated program of neurophysiological changes, which is out of our control. To give an example, if one inhales too much pepper, this brings about the automatic bodily response of sneezing. This response is not brought about voluntarily: one cannot deliberate about whether to sneeze and then sneeze. Even in cases where one realises that one is going to sneeze, it is usually very difficult to try and control this bodily response once triggered. One may sense that one is about to sneeze perhaps in a similar way to sensing the tightness of the facial muscles when one is furious and is ready to fight. The tightness of the facial muscles when angry and the opening of the mouth just before sneezing are not brought about consciously. If this is the case, although social forces can be called upon to explain the regulation of emotion, they should not be included in an

---

85 Here I use ‘automatic’ to imply that the experience of the emotion is a passive phenomenon, something that is involuntarily instigated.
86 This is not the only way to explain the automaticity of emotions. However, as I discussed in chapter 4, views which accept Basic Emotions Theory accept the claim that some emotions are hardwired. Such theories explain the automaticity as a result of the triggering of automated programs, e.g. the activation of dedicated neural circuits. See Ekman & Cordaro 2011, Scarantino 2015.
explanation of the process of the generation of emotion as this is essentially a neurobiological involuntary process.

I provide a reply against this view by arguing that emotion consists of the performance of a specific role or function that is learned rather than being hardwired. It is acquired through an interaction with the environment. With practice over large periods of time this performance may become automatic in the sense that it does not require a voluntary conscious decision so that it is performed. For example if the function of fear is to inform the organism of a near threat and to ready the organism for fight, flight or freezing, this process will be more advantageous if it happens as quickly as possible. In some cases the conscious performance of the emotion is actually internalised through learning and practice such that it becomes unconscious or automatic; such that it happens without necessitating conscious performance.87

Averill (1980) provides an account that attempts to reconcile the claim that emotion is the performance of a social role with the claim that it does not necessarily involve a conscious performance. According to Averill’s account emotions are socially constituted ‘syndromes’; they are sets of interrelated response elements including physiological changes, expressive reactions, instrumental responses and subjective experiences (Averill 1986, 100). These ‘syndromes’ are the internal representation of social norms or rules (Averill 1986, 100). Through learning the appropriate emotional roles in the developmental period these roles may become internalised – they become unconscious performances.

One way to interpret Averill’s view is to say that emotional performances work in a similar way to implicit biases. Implicit biases become integrated on the basis of repeated associations made in everyday life, experience and education/training (Sullivan-Bissett 2015). Their performance is unconscious but they are very much effective and causally

---

87 In some cases this performance is motivated by an underlying desire of the individual which one may or may not consciously call upon in every relevant situation. To give an example the emotion of anger is associated with injustice or insult suffered due to violation of a right of an individual. However feeling anger, for example when one is getting paid less for doing the same job on the basis of race, gender or ethnicity presupposes that one believes that one possesses the right to receive equal pay for doing the same job along with the desire that this right is not violated. The more situations of this type one encounters, the faster one may become in feeling anger without having to consider each time what desires their emotional behaviour is associated with.
powerful in the way that they relate to an individual’s overall psychology. To give an example, studies show that repeated associations such as the ones made between certain professions and gender (e.g. scientist and male), influence decisions in recruitment (Moss-Racusin et al. 2012); Moss-Racusin et al. (2012) found that faculty participants in an academic environment rated male applicants as significantly more competent and hireable than the (identical) female applicants. These participants also selected a higher starting salary and offered more career mentoring to male applicants. The gender of the faculty participants did not affect responses, such that female and male faculty were equally likely to exhibit bias against female students. However this influence appears to happen unconsciously in that the recruiters do not consciously decide to select an applicant on the basis of their gender category. The results of the studies show that implicit associations between gender categories and professions held by the recruiters influence the ultimate selection or rejection of applicants.\(^\text{88}\) One could argue that the performance of the emotion becomes unconscious and appears automatic in a similar way.

Averill argues that the fact that the emotional response is interpreted as an involuntary passion, as automatic rather than as a deliberate choice, facilitates its production. This facilitation is important as it enables the emotional response to happen faster, which as I claimed above is advantageous for the organism in some cases as it means that the organism will be readied for action faster.\(^\text{89}\) Averill calls this phenomenon ‘diachronic automation’ where emotional roles conform to social norms for long periods of time such that they come to play a functional role within the social system they are embedded in (Averill 1980, 336).

I agree with Averill that after long periods of learning, repetition and reinforcement from the environment, individuals need not produce emotional responses deliberately so as to conform to social models. Here it is important to note that according to the account I endorse in this thesis, emotion is identified with a cluster of several different properties such as patterns of physiological changes or the activation of neural circuits (e.g. the defence circuit in LeDoux’s research) but it cannot be reduced to any of these. As LeDoux

\(^{88}\) Several concerns have been raised against the notion of implicit bias. See Greenwald et al. 2009 for a good review of the literature on implicit bias.

\(^{89}\) This claim is easier to accept for the emotion of fear than it is for the emotion of anger. When one is afraid one’s fast response may be the difference between life and death. Whereas a fast response arising from the emotion of anger can get one into a lot of trouble and thus prove to be an unsuccessful strategy.
argues, some of these properties, e.g. the activation of defence circuits, can be found in other animals (LeDoux 2015) and could be instinctual responses to environmental stimuli, a bit like sneezing when smelling noxious substances. These instinctual responses are arguably automatic without the need for conscious or unconscious performance. It is the combination of such responses with other physiological, cognitive and phenomenological properties that constitutes the performance of emotion. In other words, in emotional performance these instinctual responses combine with other properties to create emotion.\(^\text{*90}\)

These performances become automatic through repetition, e.g. when one learns to defend oneself using martial arts without consciously having to decide upon every movement. In a similar way emotion can be performed more effectively if it becomes internalised through learning and practice. This is how Averill describes the process by which emotions are produced:

The way the components are organised into coherent syndromes is determined primarily by social and not biological evolution. Another way of stating this same idea is to say that emotions are transitory social roles – that is, institutionalized ways of interpreting and responding to particular classes of situations. (Averill 1986, 100)

As mentioned in chapter 3, Haslanger’s account of gender categories as social roles of subordination or privilege may work in exactly this way. A person can internalise their role of subordination or privilege so that they act in accordance with it. They can perform the role because it is being reinforced by the cultural environment. To give an example, the first female players of the American Baseball League had to wear dresses during games instead of the usual baseball attire. They performed this action without having internalised the belief that women need to wear dresses when performing sports. On the other hand we can think of cases where something that was enforced by society does become internalised in the performance of gender roles such as the idea that men must control the expression of emotion in public whereas women must not, or not to same degree. Or similarly the idea that it is appropriate for men to display angry or aggressive behaviour in

\(^{90}\text{As mentioned in chapter 3, none of these properties are necessary for emotion so there will be cases when these are not included in the property clusters.}\)
public but it is not appropriate for women, and so on. This suggests that people may come
to generate the emotion that they have learned is socially appropriate in that situation
without necessarily acknowledging that this is what is happening. With the passage of time
and the continuing reinforcement of this model, it may become the norm that different
gender categories are associated with the performance of different emotional roles. If this
reinforcement process happens for long periods of time, the difference in emotional
performance may be mistakenly attributed to biological factors e.g. different reproductive
organs, rather than to the social practices that brought it about. What started as a socially
enforced model may be explained through a difference based on gender or sex categories.
As in the case of emotion discussed above, because it appears to be automatic people may
believe that it is based on biological or neurobiological properties and therefore try to
explain the difference in behaviour by appealing to some biological causal factor.

Accepting the claim that emotion is socially constructed in the sense that it is constructed
in tandem by the individual and the society, does not necessarily mean that individuals
consciously choose to conform to socially constructed emotion models. Instead emotion
as a social role is reinforced by culture and society during long periods of time. Griffiths
explains this process by mentioning that emotion models are created diachronically and
not necessarily synchronically. He states that:

The existence of the model in the culture helped shape what is now a
relatively automatic reaction to certain situations. We might call this the
reinforcement version of the social role model. (Griffiths 1997, 149)

If the above claims are true then the fact that emotion appears to be automatic is
compatible with the view that it is identified with the performance of a specific function
or a role. This automaticity is a result of diachronic reinforcement of specific emotion
roles and emotional performances by society. The performance of emotion is not
necessarily conscious because it becomes internalised through learning and repetition.
Therefore, diachronic automation may explain why some emotions appear to be
automatic; why they appear to be ‘passions’ to use Descartes’ term (1648, 257) rather than
actions.
In this section I argued that the environment is involved in both emotion generation and emotion regulation. Although the dominant view considers these as two separate and consecutive processes, I argued that we have evidence against this view. I suggested that emotion regulation and generation can be considered as a single process and that the interaction with environment is an important feature within the emotion property cluster. Although emotion is associated with social properties such as interactivity and change through learning or reinforcement, this fact is not incompatible with the observation that some emotions are associated with automatic reactions. It is exactly this interactivity which may explain why some emotional reactions appear to be automatic whilst at the same time vary between historical times or different societies.

This interaction between emotion and the environment, which is crucial for the development of emotion, is a very important feature of the emotion kind which can be used to explain some of the physical properties associated with the realisation of emotional phenomena. To conclude, social properties associated with emotion such as the interactivity of emotion add support to the claim that emotion is a social kind that can be explained by looking at the functional level of psychological explanation. In the next section I discuss the claim that emotion is a natural kind, that it is a category showing an aptness for induction and explanation.

5.3 Emotion as a natural kind

5.3.1 Aptness for Induction

Identifying emotion as a social kind entails that the explanation and induction potential of the category will be tested at the functional level and not at the physical or implementational level. If emotion is a social kind such explanation and induction should be available for the full range of emotions, e.g. regret, relief, surprise, boredom, annoyance, anger and joy, as the claim here is that all of these are members of the same natural kind, emotion. In this section I give examples of successful induction and explanation to support my claim that emotion is a natural kind. The following examples of

---

91 Emotion could be considered a natural kind in several domains within the social sphere such as sociology, social ecology or social psychology. The focus on this thesis is not on identifying which social domains consider emotion a natural kind. Instead the focus is on arguing that emotion is a social kind and as such cannot be a natural kind within a domain that does not investigate the social properties associated with emotion such as its interactivity.
induction are associated with different scientific domains rather than one particular scientific domain. However what is important for my claim in this chapter is that all of these domains focus on the interaction between individuals and the social sphere and highlight the interactive and adaptive nature of emotion. This induction is the result of numerous individual observations undertaken on human and non-human animals:

- Emotion engages attention to particular aspects of the environment; it also regulates the attention span by controlling how long an individual will remain alert to some aspect of the environment (Turner & Stets 2005, 264). This induction is related to the view of early phenomenologists (Husserl 1931; Heidegger 1953; Merleau-Ponty 1945) according to which emotion is the channel through which we appoint significance to objects in the environment. By engaging attention emotion enables us to direct ourselves to specific aspects of our environment, to focus on specific objects whilst ignoring others.

- Emotion enhances retention and recollection of memory; studies show that it is easier to retain and recall emotional stimuli in our memory (Ack-Baraly et al. 2016). Memory seems to retain emotionally laden stimuli and ignore non-emotionally laden stimuli. It would not be advantageous for the organism to treat a stimulus encountered in the past as a new stimulus that needs to be evaluated anew. Instead emotion appoints a specific value to the stimulus and it is retained in memory as evaluated. This enhances the speed by which the organism will come up with the best strategy each time the stimulus is encountered.

- Emotion allows individuals to situate themselves as objects in relation to others, the situation and the environment (Colombetti 2014). This induction is associated with the interactivity feature of emotion. Through emotion we are able to engage with the environment by assigning value to certain objects or situations, e.g. judging the dog dangerous, that comment insulting etc. During that interaction we transform from subjects to objects as the environment engages us in significantly different ways, e.g. the dog threatens me, the loud noise surprises me, etc.

- Emotion can be used as a communicative means, enabling and facilitating our interaction with others. This is done via communicating our own emotional states but also by recognising and empathising with emotional states of others (Fridlund 1994). Achieving our goals in social environments presupposes that we are able both to communicate and to recognise in others things that matter to us. For
example by showing my anger to you when you insult me without provocation, I am communicating that it is significant to me that you remove the insult. When I recognise fear in the face of someone else I get information about some kind of threatening object in the environment.

- Emotion is motivating; it prompts us to some action that is usually associated with the achievement of a particular goal (Greenwood 2015; Lang 2010). If emotion enables us to navigate the environment successfully, such motivating powers are essential. It is because emotion links us to the environment that it makes sense that it motivates us to act. Emotion reveals to us things we care about and at the same time in combination with the societies we are brought up into, it provides us with strategies so as to achieve the things we care about.

- Emotion attaches to cultural norms and practices in ways that uphold the cohesiveness and maintain the structure of society (Turner & Stets 2005, 264). It seems that we maintain some norms or practices because they are associated with specific emotions or with our ability to feel emotion. For example, the practice of drama festivals in 5BC Athens, or the institution of the World Cup in football or basketball. It is because these practices are used as places of emotional generation and expression that perhaps they have been maintained for hundreds of years. These are places where emotion models are reinforced, e.g. by providing narratives for when someone should feel grief (e.g. at death of a loved one), how should one express grief (pulling out own hair, crying, feeling depressed etc.), how much pride should a person feel for one’s actions, when should someone feel enraged and what actions should one take out of anger, etc.

- Emotion malfunction correlates with an inability to form close personal relationships and with antisocial behaviour (Fantini-Hauwell et al. 2012). Evidence from social psychology shows that emotional disorders interfere with the way one interacts with the social environment (Timoney & Holder 2013). This is to be expected if the primary function of emotion is to enable us to interact successfully with the environment. Emotional disorders such as alexithymia are associated with a severe inability to form close relationships and are frequently co-occurring with major depressive disorders or psychopathic tendencies. If emotion is the channel through which we form meaningful personal relationships this is exactly what we would expect to happen. Importantly in patients diagnosed with alexithymia, such disorders can be treated in some cases through cognitive
behaviour therapy and emotional learning and training. This provides support to the claim that the social properties such as interactivity are essential for the construction of emotion.

- Emotion is open to development and change both during the developmental period and also in adulthood (De Castella et al. 2015). A study conducted by De Castella et al. (2015) shows that the regulation of emotion is connected to implicit or explicit beliefs held on whether emotion is hardwired and cognitively impetrenable and as such out of our control. According to this study, sufferers of SAD (Social Anxiety Disorder) who believe that emotion is something out of our control do not improve after CBT (Cognitive Behaviour Therapy). On the other hand, patients of SAD who believe that emotion is not out of our control show improvement or are completely cured after CBT. This may provide evidence for the interactive nature of emotion. If emotion is not a pre-programmed response and it can be regulated by beliefs held by the individual on its nature, then this fact provides support to the social nature of emotion. These beliefs held on the nature of emotion are usually a result of learning and reinforcement from the environment. If an individual learns during development that anger is an automatic reaction out of her control she is likely to believe that she has no control over it.

- Human infants develop into emotionally blunted children and adults if denied the appropriate socioemotional developmental context (Greenwood 2015, 85). As above, this provides support to the claim that emotion is an interactive kind which heavily depends on the socioemotional environment one is brought up into.

The examples given above are based on inductive inference. From observing individual cases of people suffering from emotional disorders such as alexithymia we find out that these patients show an inability to form close personal relationships. From these individual cases we infer that many people who suffer from emotional disorders are not able to form close personal relationships with others. From observing sufferers of SAD we find out that beliefs held on the nature of emotion influence our ability to regulate emotion. We then infer from the individual cases that beliefs on the nature of emotion influence our ability to regulate emotion. In a similar way, from observing individual cases of children suffering from emotional disorders we find out that these children did not have the appropriate socioemotional context during the development period. From these
individual cases we infer that children who do not have the appropriate socioemotional developmental context are likely to suffer from emotional disorders.

Such successful induction provides support for the claim that emotion is a social kind and that emotional phenomena need to be analysed at the functional level. Accepting the claim that emotion is a social kind does not, however, entail that discovering details of how emotion is physically realised does not provide additional understanding. A complete understanding and explanation of emotional phenomena will include an analysis of both the functional and the physical levels. However the analysis given at the physical level is complementary to the analysis given at the functional level; it cannot substitute it.

5.3.2 Aptness for Explanation

Our beliefs about the natural kind status of the category and more specifically our beliefs on whether emotion is a kind in some social domain have several consequences for the explanatory power of the category. As mentioned briefly in chapter 2, we are able to give successful explanations of certain phenomena by referring to natural kinds (viewed as stable property clusters) and the properties that cluster together with stability. To give an example from observing that pieces of silver have a melting point of 961.72°C we infer that silver has a melting point of 961.72°C. We can then refer to this property to explain certain phenomena e.g. Why did Marta’s knife melt? Because it was made from silver and it fell into lava with a temperature of 1100°C. Therefore we are able to provide explanations of phenomena relating to natural kinds by using the inductive inferences generated by observation of individual cases.

According to the traditional view of natural kinds (the one that I call the metaphysical view), natural kinds necessarily reveal laws of nature. The explanation of phenomena associated with them is grounded by an exceptionless law of nature e.g. in the case of silver mentioned above, it is a law of nature that silver has a melting point of 961.72°C. This is equivalent to saying that no piece of silver has a melting point different to 961.72°C. In chapter two I argued that the metaphysical view cannot accommodate many kinds in scientific domains, which show an aptness for successful induction and explanation.
regardless of whether they refer to exceptionless laws of nature. I claimed that the appeal to laws of nature should not be used as the basis that explains why natural kinds show an aptness for induction and explanation because such aptness is shown by kinds who do not reveal exceptionless laws of nature. Instead I argued that this aptness can be explained by the stability manifested by the properties in the cluster. It is because these properties are stably co-instantiated that we are able to generate successful induction for them. To give an example, if property A (e.g. feeling sad) and property B (e.g. crying) are co-instantiated with stability, one can use the existence of property A to predict the existence of property B.  

In this chapter I claimed that emotion is a social natural kind. In a similar way to the kind silver, we should be able to create successful explanation by referring to induction such as the one mentioned in the previous section e.g. by using inferences generated by the stable co-instantiation of properties in the emotion property cluster. To give a few examples, one may ask ‘Why is Stella sad?’. If emotion is a social kind that arises as an interaction between the individual and the environment, to give an explanation of this phenomenon one needs to know the context it arises in. As it happens Stella just found out that her flight is cancelled and she feels sadness because this means she can no longer attend the conference or get any of her flight money back. For this explanation to be successful one tacitly assumes a co-instantiation between the property of suffering a loss (missing the conference, losing some money) and the property of feeling sad. Such an explanation of this emotional phenomenon is based on the inference that usually when Stella suffers a loss (property A), Stella feels sadness (property B). Therefore we are able to generate successful explanation by using inductive inferences that are based on observation of emotional phenomena.

In chapter 2 I argued that the possible existence of causal factors responsible for the stability observed in the cluster properties provides additional information but should not be used to determine membership into the kind. In a similar way I claim that the model of explanation we should use for explaining emotional phenomena does not need to be

---

92 Stable co-instantiation can be used to predict the existence of property B from observing the existence of property A and also the existence of property A from observing property B. As such stable co-instantiation by itself may not be able to settle debates as to what if any, is the causal relation between properties A and B. For example James (1884) argues that it is because of property B (crying) that we have property A (Stella feeling sad).
causal; causal factors influencing the stability of the properties may not have a privileged role in explanation (Van Fraassen 1980). According to the account I suggest in this thesis emotion is an interactive kind, and the ability to instantiate it needs to be developed in a social environment; explanations of emotional phenomena will therefore be relative to the context in which they arise and the different social environments involved. For example, there may or may not be a single causal explanation available for the fact that the Japanese experience amae whereas we do not have evidence that this emotion is experienced outside Japanese society. However the absence of a causal explanation does not entail that we will not be able to create successful induction and explanation.

A few examples of what emotion-involving explanation may look like are the following:

- Some individuals kick objects out of anger because that is what they were brought up to think is an appropriate response to some situation.
- Alexithymic individuals show an inability to form close personal relationships because their capacity to experience, show or recognise emotion in themselves and others is diminished (Taylor et al. 2016).
- Religious Christians do not normally experience the emotion of envy because it is included in the seven deadly sins.
- South East Asians experience amok in certain situations within specific environments because they are brought up to think that it is the socially appropriate response to some situations.
- Patients with emotional disorders such as Social Anxiety Disorder can recover after Cognitive Behavioural Therapy because emotion is not a cognitively impenetrable primitive module (De Castella et al. 2015).
- Studies show that some adult men exhibit restrictive emotionality that is restrictions on the development and expression of emotions as a result of lack of practice during the early socialization period (Jansz 2000, 180). As Jansz claims, this restrictive emotionality is linked inextricably to the construction of masculine identities in traditional families. Such restrictive emotionality was not observed to adult men who were brought up in less traditional ways. This is possible because the development of emotion is not directly causally related to the biological feature of sex.
The above explanations are based on the assumption that emotion is a social kind: it is constructed in tandem between the individual and the social environment. Therefore the explanations are based on the assumption that experiencing certain emotions depends on societal norms and rules. This type of explanation is based on induction (e.g. in most observed cases properties A and B are co-instantiated) rather than deduction (properties A and B are always co-instantiated) which would be required if we were appealing to exceptionless laws of nature. In the example of the emotion of amok, what is implied is that in most observed cases the experience of the emotion of amok depends on societal norms and rules; if this is the case, it is highly probable that South East Asians do not experience amok in western societies because the experience of amok is not condoned in western societies.

Since there are several social domains that study emotional phenomena, there may be more than one explanation available depending on which domain one focuses on. For example there may be no single or clear explanation (causal or non-causal) for the fact that some of the British voters who opted to vote in 2016 so that the UK leaves the European Union voted in anger. However, we can provide different types of explanation by looking at the wider context/contexts e.g. socioeconomical, political, historical, etc. and offer several plausible or relevant answers. For example some of the voters were angry at the prolonged austerity measures imposed by the government but failed to direct their anger towards the government and directed it mistakenly towards the EU. Other voters may have felt anger against established policies of the EU that are considered to widen the gap between affluent classes and the majority of the population and considered the referendum an opportunity to show their disagreement with such policies. Importantly, according to the account of emotion I endorse in this thesis, for such explanations to be plausible and relevant one needs to take into consideration the social context that these emotional phenomena arise in.

One could argue that if one wishes to identify the cause of the vote one could just point to the neurophysiological state of each individual that was the cause of the physical behaviour of voting. Any other explanations we give that include references to the mental state of the voters and the context in which they arose are not actually the causes of this behaviour. My reply here would be to point out that the mental state the voters were in may not be the cause of the physical effect of voting, however, it provides information of
the causal history of this effect and, as such, it can provide an explanation of this event. To say that Jack voted so that the UK leaves the European Union because he was angry, provides a perfectly good common sense explanation of the reason he so voted.\textsuperscript{93} Any information concerning the individual neurophysiology of the voters can shed additional light but it is a different type of explanation.\textsuperscript{94}

5.4 Emotion as a social natural kind

The claim that emotion is a social natural kind forms the conclusion of an argument that includes the main claims established in chapters 2, 3, 4 and 5 as premises. As mentioned in chapter 1 of this thesis, the argument has the following form:

P1- Natural kinds are property clusters that show an aptness for successful induction and explanation (epistemological view – chapter 2) regardless of the possession of similar causal mechanisms or the inclusion of interactivity or normativity properties (chapter 3).

P2- Some social kinds can be considered natural kinds according to the epistemological view because they show an aptness for successful induction and explanation (chapter 3).

P3- Emotion is a social kind because the emotion property cluster includes social properties (chapter 5) that are necessary for the explanation of emotional phenomena. These properties cannot be reduced to physical or neurobiological properties (chapter 4).

P4- Emotion is a social kind which shows an aptness for successful induction and explanation (chapter 5).

\[ C \rightarrow \text{Emotion is a social natural kind.} \]

In chapter 2 (see Premise 1 above) I argue that natural kinds are kinds that permit successful generalisation and induction according to the epistemological view. According

\textsuperscript{93} This worry relates to a version of the causal exclusion problem attaching to mental properties when these are not considered reducible to physical properties (Kim 2000). Here I am borrowing from Beebee (2017) to claim that mental properties such as Jack’s anger in the example above, are explanatorily powerful regardless of whether they are causally relevant to Jack’s voting one way or another.

\textsuperscript{94} Here I leave out any pathological cases because in such cases emotional processes may require completely different types of explanation.
to this view, natural kinds are property clusters where none of the properties is necessary or sufficient for membership. As discussed in chapter 2, according to cluster accounts of natural kinds, none of the properties included in the cluster is considered necessary or important for membership. However, this does not entail that the property cluster is not stable enough to guarantee successful induction and explanation. These properties reliably and stably occur together.

As none of the clustered properties are necessary or sufficient for membership, in some cases, some of the properties may be missing. For example, in some cases the phenomenological aspect may be missing; alternatively there may be cases where emotion does not lead to the appropriate motivational states, or where we experience bodily feelings without connecting them to an object in our environment. In such cases one can experience bodily states normally associated with emotion such as increased heart rate or skin conductivity, without actually integrating these bodily states with stimuli provided by the environment, thus lacking intentionality. In these cases it is likely that one will not appoint any significance to objects in the environment such that one is not prompted to a specific action and as a result one will not be motivated to act one way or another. Although such cases make it difficult to distinguish emotional phenomena from non-emotional phenomena, the account I endorse allows for cases of uncertainty relating to the assumption that natural kinds do not necessarily have clear and distinct boundaries (Boyd 1991, 142; 1999, 153). As Boyd states, accepting that such cases exist should not imply that we should ignore the numerous cases where there is no uncertainty and the properties cluster in a stable and reliable way (1991, 141).

According to the account of natural kinds I propose in chapter 3, all natural kinds are socially constructed kinds because they are constructed by and dependent on the interests, purposes and theoretical methodologies of several disciplines (Boyd 1991, 143). Some natural kinds refer to social phenomena, that is, phenomena that need to be investigated within the social domain. In chapter 3 (see Premise 2 above) I argue that social kinds can be considered natural kinds on the basis that they show an aptness for successful generalisation and induction and that some social kinds show this aptness regardless of whether they share similar underlying causal mechanisms and whether their clusters include interactive or normative properties. In chapter 3 I focus on the interactive properties and heterogeneous construction, as these features are used by several
philosophers (Hacking 1999; Griffiths 1997) as evidence for the claim that some social kinds are not natural kinds. There are several mechanisms responsible for the stability of the clustered properties, e.g. natural selection, the environment one is living in, the society and culture one is brought up into, etc. However, as discussed in chapter 3, the fact that there are several mechanisms underlying the stability of the clustered properties does not entail that the properties do not cluster with stability. In addition and, as discussed in chapter 3, although some clustered properties are associated with different causal factors, the natural kind status of the cluster does not track identical causal mechanisms. It specifically tracks the stability of the clustered properties and is based on whether this stability is sufficient for enabling successful induction and explanation.

There are several different theories on the nature of emotion, however according to the most dominant ones emotion is either a neurobiological kind (e.g. BET) and its property cluster consists of neurobiological properties, or it is a kind in social psychology or social ecology and its property cluster includes properties best explained within social domains (Averill 1986; Griffiths 1997; Mallon 2016). In chapter 4 I argue that emotion should not be identified as a property cluster that includes strictly neurobiological properties. Researchers such as Griffiths (1997), Scarantino (2015) Russell (2015) and Barrett (2006) suggest that the category of emotion is too heterogeneous to be considered a natural kind. It includes phenomena that do not share any clear and distinctive boundaries (Russell 2015, 204), or any similar underlying causal mechanisms (Griffiths 1997). These researchers suggest that the term ‘emotion’ is not a scientific term and it should be eliminated from scientific vocabularies or retained simply as an umbrella term. For example Russell calls it ‘a chapter heading which does not function as a technical term’ (2015, 205). I argue in chapters 2 and 3 of this thesis that it is not necessary for emotion to have clear or distinctive boundaries or similar underlying causal mechanisms for it to be considered a natural kind. The objection from heterogeneity seems to be based on the problem of variability raised against Basic Emotion Theory. As mentioned in chapter 4, empirical data provide evidence against correspondences between specific emotions and neurophysiological responses. As discussed in chapter 4, this would only affect the natural kind status of the emotion if emotion was a kind in neurobiology and the hypothesis associated with this claim was that we will be able to discover regularities at that level of analysis.
A big area of disagreement between social constructionist and basic emotion theories is whether the neurobiological properties of emotion have any explanatory power for emotional phenomena (Griffiths 1997, 167). I argue in chapter 4 that the emotion property cluster includes neurobiological properties but that these properties are contingent. This does not mean that they do not provide additional information as to how emotional phenomena are physically realised. However on their own they cannot provide an adequate explanation of emotional phenomena. To give an example, if we wish to provide an explanation for the fact that Sally is angry at John we will need to know more than the physiological changes and the neurophysiological activation of Sally’s body. We need to know general information about Sally’s health, personality, preferences and interests as well as the larger social context that Sally is living in, e.g. whether or not Sally’s anger is deemed an appropriate emotion for Sally vis a vis the situation, her relationship with John, her social status, etc.

In chapter 1 I argue that our theories on the natural kind status of emotion influence the design of experiments and the interpretation of empirical data. In this chapter I argue that our beliefs on the nature and natural kind status of emotion influence our attitudes towards emotional phenomena and any other social or psychological phenomena associated with them. I argue that our beliefs on the nature of kinds such as gender categories or emotion are important because they can distort the actual relations between our mental states and our behaviour. We may mistakenly assume causal relations where there are none and give erroneous explanations of mental phenomena. This is because according to a popular and, as I claim in this thesis, mistaken view, emotions are automatic reactions caused by pre-set neurobiological properties. If emotion were a neurobiological kind, it would be easier to accept that we have no control over it in the same sense that we (almost always) cannot control bodily reflexes such as itches or hiccups. I claimed in this chapter that the feature of automaticity shown by some emotions can be explained within an account of emotion as a social kind.

In this chapter I also argue that emotion shows an aptness for induction and explanation at the functional level. Of course we may be able to discover regularities in the implementational level if we narrow the focus to the physical systems that realise emotion, for example by restricting the investigation to specific species or genera. For example mammalian emotion may have more similarities at this level than emotion realised in
mammals and molluscs. Therefore the heterogeneity shown by properties realising the function of emotion is not caused by the fact that emotion refers to different things for which we cannot discover successful induction and explanation. Instead it is caused by the fact that emotion is a social kind that is realised by different physical properties. In addition, the heterogeneity of causal factors is intimately connected with the feature of interactivity. As emotion is a complex and evolving adaptation which needs to be developed through interaction with the environment, its property cluster allows for variation and change from both the environment and the individual. If one accepts the premises argued in the chapters of this thesis, it follows that emotion is a natural kind notwithstanding the fact that it is a social interactive kind.

5.5 Implications for the study of affective phenomena

In what follows I argue that changing our attitudes towards the nature and explanation of emotional phenomena has important consequences in all areas which study affective phenomena, e.g. social and experimental psychology, sociology, gender and cultural studies, clinical psychology, neuroscience, psychiatry and law. Approaches in neuroscience or experimental psychology that identify emotion with a single component do not provide an accurate analysis of emotional phenomena.

Herschbach and Bechtel argue that the move from what they call ‘natural kinds’ theories of psychological phenomena to the psychological construction theories is the transition ‘from simple localization to more complex and dynamic accounts of the neural mechanisms responsible for these psychological phenomena and potentially to a reconceptualization or reconstitution of mental phenomena themselves’ (Herschbach & Bechtel 2015, 41). It seems that Herschbach and Bechtel group together all theories that claim that emotion is a natural kind with BET or with theories that postulate localised distinct brain regions for each individual emotion. As I claim in chapters 2 and 4, however, the epistemological account of natural kinds rejects such claims. Emotion members do not share necessary and sufficient properties on the basis of which we determine membership. As such, my account differs from componential accounts such as the one endorsed by Herschback and Bechtel (2015), and psychological constructionist views such

---

95 This applies to the category of emotion and to individual categories of emotion such as fear or anger.
as the ones advocated by Barrett (2015), because these accounts conclude that if BET is false, *emotion* is not a natural kind. Instead I maintain that *emotion* is a natural kind according to the epistemological view. We can create successful induction and explanation relating to *emotion* if *emotion* is considered a kind in social psychology or some other social domain.

I mentioned in 1.2 that implicit or explicit beliefs held on the natural kind status of *emotion* influence the design of experiments and the interpretation of results in affective science. Accepting that *emotion* is a social kind entails that providing accurate explanations of emotional phenomena would require a wider approach when interpreting results of experiments. To give an example, if *emotion* is a property cluster without any necessary or sufficient properties, experimenters should use a wider range of methods to create generalisations between emotions, physiology and behaviour that take into account the wider context of emotional phenomena, which may include the socioeconomical background or history of the participants. In addition when designing experiments experimenters should no longer aim to discover an ‘essence’, viewed as necessary or sufficient properties or identical underlying causal mechanisms, for each emotion by monitoring exclusively the neurophysiological activity of participants. Here I do not claim that such research is of no value. On the contrary: such research is important and necessary for revealing the physiological properties of emotion.

Another important consequence of the view that *emotion* is a social kind relates to the explanation of emotional phenomena. If one accepts the claim that *emotion* is a social kind, one should reject the claim that *emotion* is a neurobiological phenomenon that once activated it cannot be stopped; that it just happens to us without our control. For example, one may kick the table in anger and erroneously explain this act as arising from the neurobiological kind of *anger*. If one believes that emotion is a neurobiological mechanism which, once activated, is out of our control, one may also believe that the action of ‘kicking the table’ arose as a result of an activation of an automated programme e.g. *anger* which is out of one’s control.96 Therefore if one has erroneous beliefs about the nature of *emotion* one may give erroneous explanations of the relation between one’s mental states and corresponding behaviour. This mismatch can also occur in cases where socially

---

96 This is not to say that there is no neurobiological or physiological ingredient in an emotion. Instead as I argued in the previous chapter the physiological aspect should not be identified with emotion. Emotion is not located at the biological level, instead it requires the combination of further higher cognitive ingredients.
constituted roles become internalised. However in those cases, as soon as one becomes aware that these roles are socially constructed one can decide to stop practicing them. This is not to say that this is an easy process. According to my account these performances are a result of long periods of practice, learning and interaction with the environment. Therefore the act of ceasing to perform them is likely to be a lengthy process. This view has consequences for the diagnosis and treatment of several emotional disorders such as alexithymia.

Having mistaken beliefs about the nature of emotion and correspondingly about whether and in what way emotional performance can be controlled has consequences for the disciplines of sociology, anthropology, gender and cultural studies. Another way to describe this phenomenon is to say that emotion is a covert (hidden) construction. Although people may believe that emotion is a neurobiological category, emotion is actually a social construction, it is a social kind; it is continuously constructed in tandem by each individual and the society the individual dwells in. It is a covert construction because according to a popular view, people erroneously consider it to be a neurobiological category. As soon as people find out its true nature, this disrupts the process by which the category is constructed (Griffiths 1997, 147).

For example one may have the mistaken belief that emotion is a neurobiological mechanism and emotions are hardwired programs of responses which, once activated, are out of our control. As long as one believes that this is true, although one may attempt to control actions out of emotion, one does not attempt to control the emotion itself because one does not believe that such control is possible. This means that the beliefs we hold on the nature of emotion, influence our ability to control and explain emotion and action out of emotion. If emotions are identified with the activation of automated systems in the brain, attitudes we take toward them could not influence or control this activation. Importantly this ability to exert control is not limited to controlling our actions out of emotion. It also includes the physiological changes and the phenomenology of emotional experience. It therefore has an impact on emotional experience in a holistic way.

97 Here I am making a weak claim that one finds it easier to change emotional performance or behaviour when this performance is not associated with some kind of biological/physical mechanism.
As mentioned in the previous section, clinical studies in psychology provide support for the claim that emotion (generation and regulation) is not out of our control (De Castella et al. 2015). One explanation for this finding is that the inability to access the mental states that actually cause and explain certain behaviour is revealed by CBT to patients such that, as soon as they discover that emotion is not an involuntary phenomenon, they are able to control and regulate it. By acknowledging that emotion (and emotions) is not a hardwired automatic reaction that is out of our control, implicit beliefs held by people about the nature of emotion are removed; as such people can have access via CBT to the real causes behind emotional behaviour and emotional disorders. Mallon in a similar way argues that we construct causal explanations of our mental states on the basis of our background theories about what can and cannot constitute a plausible cause for them (Mallon 2016, 125). As a result, we may decide that our mental states have completely different causes than the ones they actually do. He states:

The result is that we might decide, along with others in our epistemic community, that our mental states and behaviours have completely different causes than they in fact do….if subjects don’t realise that the desire to act out multiple personality disorder (MPD) symptoms, or, say, act out racial or gender or emotion templates, is instrumental to the satisfaction of other desires, then desires on which they instrumentally act may seem simply like primitive desires that subjects have as part of their natures. (Mallon 2016, 125)

To give an example, Alex may experience anger when he is challenged during a debate because he was brought up to believe that in such situations this emotion is appropriate for people of his race (white) or gender (man). Here the desire to act according to the norms or rules associated with certain groups because one identifies with such groups is the motivating force behind this behaviour. However, what happens with covert constructions is that Alex may erroneously believe that his experiencing anger in this situation is something that is associated with him biologically, as part of his nature (e.g. being a white male) instead of something that he has learned to do. Therefore having inaccurate beliefs on the nature of emotion causes us to provide false explanations of emotional phenomena and actions out of emotion. This misrepresentation of emotion and its relation to biology may contribute to the creation and maintenance of gender or racial stereotypes.
In addition, having accurate beliefs about the nature of emotion has consequences for appointing agential responsibility in law environments. If we think that the causes of our emotional behaviour are not under our control, this partially removes agential responsibility for actions out of emotion. Mallon claims that this has a very surprising result for our capacity to act as responsible agents:

That the capacity to act as a free agent with regard to actions of a certain type depends upon shared understanding of the sort of mechanism from which the action emerges. Widespread false accounts of a behaviour-producing mechanism (of the sort that social constructionists supposed to obtain in a community) in fact genuinely undermine our capacity for responsible agency. (Mallon 2016, 133)

According to the passage quoted above, if we do not know what an emotion is and we hold inaccurate beliefs about what causes emotion, we will also create inaccurate explanations for the behaviour of agents when it is somehow related to emotion. In some cases this may lead to removing the responsibility for an action from the agent and explaining the action as some involuntary neurobiological process such as sneezing. For example in several Western societies one is likely to get a more lenient sentence for killing someone when enraged instead of killing someone *in cold blood*; if the jealous husband murders the wife because he thought she was being unfaithful while enraged, the fact that he was under the hold of an emotion may result to a more lenient penalty in many Western societies. If, as I argue in this thesis, emotional action is importantly influenced and depends on the social environment for its generation, such practices of associating action out of emotion with lenient penalties must be revised.
Conclusion

In this thesis I put forward a claim that *emotion* is a social natural kind. I argued that emotion is what enables us to successfully navigate social environments. It is an interactive kind that is evolving to match changes in the environments we find ourselves in. Its function can be realised by different physical systems. It needs to be developed through learning so that it enables us to interact with our environment successfully and form close and meaningful relationships. It includes within its cluster social properties such as interactivity and normativity. I argued in chapter 3 that the existence of such social properties within the property cluster of emotion does not entail that *emotion* cannot be considered a natural kind. What determines the natural kind status of the category is the aptness shown for successful induction and explanation.

In chapter 4 I claimed that we may not find regularities in the implementational level because *emotion* can be multiply realised at that level. Instead, as shown in the second part of chapter 5, successful explanation and induction are possible at the higher level of explanation, namely the functional level, which analyses the function *emotion* is performing. Changing our beliefs about the nature of emotional phenomena and the natural kind status of *emotion* has implications for all disciplines involved in the scientific study of emotion, e.g. social psychology, sociobiology, sociology, psychiatry and affective computing. In the last section of this thesis I suggest how the change in beliefs can generate progress and lead to a better explanation of emotional phenomena.

In chapter 1 I discussed three dominant views on the natural kind status of *emotion*; according to view A, some emotions (the ones called ‘basic’) are natural kinds but the category of emotion is not a natural kind (Ekman 1972a; Griffiths 1997; Scarantino 2015). According to view B, all emotions are basic emotions or combinations of basic and non-basic emotions and the category of *emotion* is a natural kind (Prinz 1994; Charland 2004). According to view C, basic emotions are not natural kinds and the category of *emotion* is not a natural kind (Barrett 2006; Lindquist et al. 2013). I argued against the first two views by putting forward two claims. First I argued that basic emotions are not emotions (chapter 4). Second I argued that *emotion* is not a neurobiological kind but a social kind (chapter 5). Although I agree with view C that we have reasons to doubt that basic
emotions are natural kinds, I disagree with the claim that *emotion* as a category is not a natural kind. Instead, I claim in chapter 5 that *emotion* is a social kind because its analysis and explanation involves important social properties. In addition, I argued that *emotion* shows aptness for successful induction and explanation and, as such, that it meets the criteria for natural kind status. In short, *emotion* is a social natural kind.
Bibliography


Bonta, B. 2017. ‘The Piaora and the Giant Spiders’. Article on *Peaceful Societies* website, University of Alabama at Birmingham, College of Arts and Sciences, Department of Anthropology https://cas.uab.edu/peacfulsocieties/2017/03/02/piaora-giant-spiders/


James, W. 1884. ‘What is an emotion?’, Mind, 9: 188-205.


