EXPLICATING EMOTIONS

by

Andrea Scarantino

B.S., Economics, Bocconi University, 1994
M.S., Philosophy of the Social Sciences, London School of Economics and Political Science, 1997
Ph.D., Economics, Universita' Cattolica, 2000
M.A., Philosophy, University of Pittsburgh, 2005

Submitted to the Graduate Faculty of

University of Pittsburgh in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2005

UNIVERSITY OF PITTSBURGH

FACULTY OF ARTS AND SCIENCES

This dissertation was presented

by

Andrea Scarantino

It was defended on

July 20, 2005

and approved by

Paul Griffiths, ARC Federation Fellow and Professor of Philosophy, Department of Philosophy, University of Queensland (Co-Director)

Peter Machamer, Professor of Philosophy, Department of History and Philosophy of Science, University of Pittsburgh (Co-Director)

Bob Brandom, Distinguished Service Professor of Philosophy, Department of Philosophy, University of Pittsburgh

Ruth Millikan, Emeritus Professor of Philosophy, Department of Philosophy, University of Connecticut
(Outside Reader)

EXPLICATING EMOTIONS

Andrea Scarantino, PhD

University of Pittsburgh, 2005

In the course of their long intellectual history, emotions have been identified with items as diverse as perceptions of bodily changes (feeling tradition), judgments (cognitivist tradition), behavioral predispositions (behaviorist tradition), biologically based solutions to fundamental life tasks (evolutionary tradition), and culturally specific social artifacts (social constructionist tradition). The first objective of my work is to put some order in the mare magnum of theories of emotions. I taxonomize them into families and explore the historical origin and current credentials of the arguments and intuitions supporting them. I then evaluate the methodology of past and present emotion theory, defending a bleak conclusion: a great many emotion theorists ask “What is an emotion?” without a clear understanding of what counts as getting the answer right. I argue that there are two ways of getting the answer right. One is to capture the conditions of application of the folk term "emotion" in ordinary language (Folk Emotion Project), and the other is to formulate a fruitful explication of it (Explicating Emotion Project). Once we get clear on the desiderata of these two projects, we realize that several long-running debates in emotion theory are motivated by methodological confusions. The constructive part of my work is devoted to formulating a new explication of emotion suitable for the theoretical purposes of scientific psychology. At the heart of the Urgency Management System (UMS) theory of emotions I propose is the idea that an “umotion” is a special type of superordinate system which instantiates and manages an urgent action tendency by coordinating the operation of a cluster of cognitive, perceptual and motoric subsystems. Crucially, such superordinate system has a proper function

by virtue of which it acquires a special kind of intentionality I call pragmatic. I argue that "umotion" is sufficiently similar in use to "emotion" to count as explicating it, it has precise rules of application, and it accommodates a number of central and widely shared intuitions about the emotions. My hope is that future emotion research will demonstrate the heuristic fruitfulness of the "umotion" concept for the sciences of mind.

TABLE OF CONTENTS

LIST OF FIGURES

PREFACE

I decided to devote my dissertation to the study of the emotions when Paul Griffiths joined the History and Philosophy of Science Department at the University of Pittsburgh. Initially, my idea was to use the emotions as a Trojan horse to enter the citadel of morality. I soon realized that I could not find the entrance to the horse. Three years later, I am still looking for it, but I feel the journey has been worth it. The emotions have proven to be a fascinating topic in their own right, and a profoundly challenging one. I am more convinced than ever that they hold the key to understanding a number of phenomena we care deeply about, including morality, art, mental disorder and rational decision-making.

There are several people I would like to thank, starting from my co-directors, Paul Griffiths and Peter Machamer. Without Paul, this dissertation would simply not have been written. He has been a terrific and highly engaged advisor, and a true friend. I learned a great deal from his work on the emotions, and his philosophical talent, quick wit, and prodigious memory have been a continuous source of inspiration for me. I will always treasure the memory of our regular discussion meetings at the Coffee Tree in Squirrell Hill.

Peter Machamer has been a co-director in the last, and crucial, year and a half of the dissertation, offering a great deal of excellent philosophical advice, both verbally and in writing. Peter's influence is especially evident in my attempt to articulate a notion of affordances suitable for shedding light on the intentionality of emotions. Peter has been a teacher of life as well as philosophy. He has taught me by example the values of tolerance, generosity and communal living. I have been a guest at Peter and Barbara's house many times, and I have always gone home with a warm feeling of gratitude and joy. I will miss Peter and Barbara' friendship enormously.

Bob Brandom has taught the course from which I have learned the most in graduate school, namely Metaphysics and Epistemology. I highly recommend taking this course with him. I can't imagine a better way to be introduced to the foundational questions in M&E than through the secure guidance of Bob Brandom. The rigor of his philosophical thinking is what

lies at the foundation of his talent as a teacher. I am grateful for all I have learned throughout the years from Bob's courses, articles and books.

Ruth Millikan has been wonderful in her availability to me. I met her through my dearest friend Bruno Galantucci a few years back. I still remember our first philosophical conversation - on pushmi-pullyu representations - by a placid lake in Storrs, Connecticut. What was most inspiring about it was Ruth's openness towards me, then a perfect stranger. She did not make me feel like I was wasting her time, even though I am afraid I was. Her exemplary generosity has shined through in the following years. Ruth has given me some of the most detailed, probing and philosophically brilliant comments I have ever received on my work. A number of the central ideas of my dissertation have emerged from reflecting on her comments, and trying to deal with the difficulties they raised. All of this, Ruth has always done with a smile and with the utmost kindness and consideration.

I also want to thank the History and Philosophy of Science Department, which has been a great home for me, and the Philosophy Department, which has been essential to my philosophical upbringing. Many thanks also to Rita Levine and Joann McIntyre for many years of much appreciated help, and to all HPS graduate students, a wonderful group of friends and fellow travelers I will miss a lot. Finally, I want to express my deep thankfulness to my father Franco, to my mother Adriana and to my brother Davide, for a lifetime of unconditional love and support.

1. INTRODUCTION

1.1. WHAT IS AN EMOTION?

Until very recently, a common incipit for a book or article on the emotions took the form of a complaint: Why have the emotions been neglected for such a long time? The complaint was followed by a statement to the effect that this was indeed a shameful state of affairs, because the emotions mattered a lot in the theorist's field of expertise. This sort of incipit has now become anachronistic. The emotions have indisputably become an object of intense interest in a spate of disciplines. This is testified by the constant output of new conferences, handbooks, monographies, journals and articles devoted to them. The shared insight is that emotions hold the key to understanding a number of phenomena we care deeply about. Just to mention a few, the emergence of morality (Gibbard 1991, Nussbaum 2001, Haidt 2000), the perception of art (Kivy 1989, Laver and Hjort 1997, Matravers 2001), the evolution of minds (Ekman 1999b, Tooby and Cosmides 1990, 2000), rational decision-making and mental disorders (Damasio 1994, 1999), and the neurobiological bases of behavior (Le Doux 1996, Panksepp 1998).

But what is an emotion? In 1884, William James asked this very question in the title of a celebrated essay in which his theory of emotions as perceptions of bodily changes was first introduced. It seems fair to say that James' question has yet to be satisfactorily answered. This is certainly not for lack of trying. Before the relative neglect to which they were subjected in the first seventy years of the 20th century, the emotions have been an object of intellectual speculation for centuries, ever since Aristotle and the Stoics began developing complex accounts of their nature and value. The cumulative effect of these efforts is a dazzling range of answers to James' question. Just to mention a few especially popular ones, emotions have been characterized as judgments, perceptions of bodily changes, behavioral predispositions, biologically based solutions to fundamental life tasks, and culturally specific social

constructions. According to some researchers, on the other hand, emotions have nothing in common other than being designated by the term "emotion" in English. Under this view, trying to theorize about "emotion" writ large is a waste of time.

It is hard not to feel overwhelmed by the sheer variety of answers and approaches to James' question, and skeptical about the possibility of an emerging consensus. After all, the emotions have been studied since Ancient Greece, and the antagonism between competing research programs does not seem to have abated through time but, if anything, increased. To apply broadly conceived Kuhnian categories to today's debate, emotion theory can be described as being in a state of crisis. An old paradigm, embodied by the cognitivist tradition, dominated from the early 1960s to the early 1990s. According to this paradigm, emotions are essentially judgments or appraisals of a particular kind. Although well-known researchers such as Robert Solomon (2003) and Martha Nussbaum (2001) still work within this tradition, in the last twenty years the central commitments of cognitivism have all been progressively undermined. Cognitivists have been accused, persuasively in my view, of having overintellectualized the emotions, and failed to account for some of their most important phenomenological and motivational features (Griffiths 1997, Delancey 2001).

At the same time, no new paradigm has yet emerged to substitute the old one. Paul Ekman's (1999b) and Carrol Izard's (1992) affect program theory, Antonio Damasio (1994, 1999, 2003) and Jesse Prinz's (2004a, 2004b) Neo-Jamesianism, and Brian Parkinson (1995, 2005) and Paul Griffiths' (2003, 2004) transactionalism are arguably the three most influential contemporary alternatives to cognitivism. Each of them, however, faces its share of substantive objections, and it is unclear that any of the competing accounts currently on the table has the resources to overcome them.

The main objective of this dissertation is to put some order into what appears to be an intractably chaotic domain of investigation, and offer a tentative way out of the state of crisis I described by offering a novel theory I call the Urgency Management System (UMS) theory of emotions. My strategy relies on three main moves, which I carry out respectively in the Historical Part, in the Methodological Part and in the Constructive Part of my dissertation. Let us briefly consider what each part aims to achieve.

1.2. HISTORY

A careful analysis of the history of emotion theory reveals that, despite the existence of significant differences between research traditions, there is more common ground between rival theories of emotions than it may at first appear. Five main traditions, I argue, have battled for the soul of emotions in the last 2,500 years, emerging at different times in the history of the subject. I call them the *feeling tradition*, the *cognitivist tradition*, the *behaviorist tradition*, the *evolutionary tradition* and the *social constructionist tradition*. Each of them comes in many flavors, and several authors belong to more than one tradition at the same time. Each tradition can be usefully characterized in terms of a cluster of core intuitions about the emotions, and of a specific sensibility for what is theoretically interesting about them. The historical investigation I propose tries to recover the contours of an area of consensus across distinct traditions, and learn from the insights and mistakes of each tradition. Some aspects of this consensus are worth maintaining, and represent the positive legacy of centuries of investigation. Here is a short summary of a few of them.

At this stage of research, emotion theorists of all stripes agree by and large that emotions have intentionality, even though the proper characterization and explanation of such intentionality are very much up for grabs. Also, most emotion theorists agree that some emotions exist not only in adult humans but also in animals and infants. Moreover, in the last twenty five years the idea that emotions are elicited - at least some of the time - by a mechanism which is fast, mandatory, and cognitively impenetrable has gained wide currency. It is an open question whether we should call an input system with such features a module, but it is certainly the case that any good theory of emotions should account for their peculiar elicitation and relative insulation from cold-blooded reflection. Finally, emotion theorists share an understanding of what we may call the *marks of emotionality*, by which I mean the prototypical components involved in instances of prototypical emotions such as anger, fear, sadness, or disgust. Such components comprise an evaluation, a suite of physiological responses, a conscious experience, and a behavioral action tendency manifested by physical actions, mental actions and expressions.

The distinctions between rival theories of emotion emerge when it comes to putting such components together in the form of an answer to the question: “What is an emotion?”. My investigation of past and present emotion theories indicates that the attempt to answer such

question always ends up encountering the same problem, which is that there seem to be items we call emotions in ordinary language which fall outside the purview of the theory, and items which we don't call emotions which fall within it. The common reaction to this problem is to go back to the drawing board, and try to formulate an account of emotions which eliminates the existing counterexamples without encountering any new ones. But is this an interesting project? And under what conditions can it be successful? These are some of the questions the Methodological Part of my dissertation has tried to address.

1.3. METHODOLOGY

The history of emotion theory presents us with a vivid portrayal of what generations of emotion theorists have tried to achieve. Two aims stand out because of their ubiquity. The first is that of *ordinary language compatibility*. As I mentioned, emotion theorists of all ages have strived to offer an account general enough to encompass all and only those things that are called "emotion" (or "anger", "fear", "shame", etc.) in ordinary language. The second aim is that of *theoretical fruitfulness*. Emotion theorists have aimed to develop accounts of emotions which can further our understanding of their value, function, control, origin, and relation to other faculties of theoretical interest (e.g. rationality).

My central point is that the project of achieving ordinary language compatibility and the project of achieving theoretical fruitfulness should be divorced from one another. Not only do they have distinct desiderata and require a distinct methodology, but it is very unlikely that they can be fulfilled by one and the same account of emotions. This is because, as I will argue, folk emotion categories are too heterogeneous and vague to allow for anything more than a family resemblance account, which is unlikely to be theoretically fruitful. The job of articulating a family resemblance account of emotion and its subordinate categories is left to what I call the Folk Emotion Project. If an emotion theorist is instead primarily interested in theoretical fruitfulness, she should "explicate" folk emotion categories, roughly along the lines first established by Rudolf Carnap (1950) for this intellectual endeavor. I call this the Explicating Emotion Project, which is the project I am personally interested in.

Once the desiderata for these two projects are elucidated, it appears clear that some of the debates in which emotion theorists engage are based on methodological confusion. One form of confusion is to try to offer a definition of emotion in the context of the Folk Emotion Project, namely a set of individually necessary and jointly sufficient conditions for something to count as an emotion in the ordinary sense. If folk emotion categories are heterogeneous and vague as I claim, this project is bound to fail no matter how many times we try to carry it out successfully. The other form of confusion is to assume that only one definition of emotion can be offered in the context of the Explicating Emotion Project, and that it must avoid ordinary language counterexamples in order to be good.

However, when we engage in the activity of explication our fundamental objective is to endow the explicandum with fruitfulness relative to the objectives of a theory. We can provisionally understand the fruitfulness of what I will call the *explicatum* in terms of its suitability for being embedded in the classificatory, explanatory and predictive activities of a given theory. As I will argue, the explicatum of a folk emotion category only needs to achieve "similarity in use" with the explicandum. The payoff for giving up a portion of ordinary uses is to endow the explicatum with a fruitfulness lacked by the explicandum. Crucially, the same folk emotion category can give rise to many good *explicata*, each of which will be fruitful with respect to the distinct theoretical objectives of some theory T, but none of which will be insulated from all ordinary language counterexamples.

1.4. THEORY CONSTRUCTION

The objective of the Constructive Part of my dissertation is to propose a new explicative theory of emotions fruitful relative to the objectives of scientific psychology. At the heart of the Urgency Management System (UMS) theory of emotions I introduce and defend is the idea that an "umotion" is a special type of superordinate system which instantiates and manages an urgent action tendency by coordinating the operation of a cluster of cognitive, perceptual and motoric subsystems. Crucially, such superordinate system has a proper function by virtue of which it acquires a special kind of intentionality I call pragmatic. I call it "umotion" to signal that I am in

the business of explicating emotion rather than capturing all and only the meaning of emotion, and that I take *Urgency* to be the fundamental feature of the explicatum I offer.

There are two main novelties to the UMS theory. The first is a new account of the *vehicle* of emotional representation, and the second is a new account of the *representation relation* between an emotion and what the emotion is about. Consider the two most popular theories of emotion in contemporary philosophy, namely cognitivism and Neo-Jamesianism. According to cognitivism, the vehicles of emotional representation are judgments. Under this view, an emotion such as fear can be identified with the judgment that danger is present, and it represents in the same way in which judgments represent. This approach, among other things, overintellectualizes the emotions, and it fails to account for their motivational dimension. According to Neo-Jamesianism, an emotion is a perception of bodily changes. The representation relation is explained in teleosemantic terms, by arguing that what such changes represent is what they have the function of carrying information about. Under this view, an emotion such as fear is to be understood as a perception of fear-typical bodily changes with the function of being elicited by danger. The problem is that many emotions lack concomitant bodily changes, and that it is unclear what function the correlation between states of affairs and perceptions of bodily changes *as such* could possibly serve. What is missing, once again, is an appreciation of the motivational dimension of emotions, namely of the fact that emotions correlate with features of the environment and provide behavioral guidance *at the same time*.

According to the UMS theory, the vehicles of emotional representation are *urgency management systems*, i.e. umotions, namely systems which offer global coordination of organismic resources in situations which demand the pursuit of high priority goals. The classic criticism of views of this sort is that they fail to account for the intentionality of emotions, namely for the fact that emotions appear to have constitutive conditions of appropriateness.

To respond to this criticism, I provide a new theory of the *representation relation* between the emotional vehicle and what the emotion is about. I argue that emotions are *intentional pushmi-pullyu representations*, which combine descriptive and directive purposes into an undifferentiated whole. This idea is borrowed from Ruth Millikan's (2004) theory of intentionality, but applied to emotions in novel ways. Under the view I propose, an emotion such as fear is to be identified with an urgent avoidance tendency with the proper function of being elicited by danger. This approach brings the motivational dimension of emotions center stage,

and allows us to appreciate that the potential benefits generated by emotions are primarily related to the special action control structure they embody.

1.5. PLAN

Part 1, from chapter 2 to chapter 7, aims to shed light on the understanding of the emotions favored by the five main traditions which I take to have shaped the history of emotion theory. Given the extent of the period covered, I had to make some difficult choices concerning whom to focus on and whom to neglect. The overarching criterion for my choices has been the degree of influence on the emotion theorists that followed. The authors I focus on have all either significantly changed or revolutionized the history of the subject. Here is a summary of the basic tenets associated with each tradition, jointly with an example of their application and a few representative authors:[1]

[1] Although some of the representative authors belong to more than one tradition (e.g. Griffiths belongs to both the evolutionary and the social constructionist traditions), I disregard this complication for the sake of simplicity.

TRADITIONS	An emotion is essentially…	Example	Authors
Feeling tradition	… a special state of consciousness/ bodily state	Anger is a state of high and unpleasant autonomic arousal characterized by increased heart beat, blood pressure, rate of respiration, and gastric activity, decrease in saliva flow, trembling, etc.	Aristotle, R. Descartes, D. Hume, S. Freud, W. James, A. Damasio, J. Prinz
Behaviorist tradition	…a special disposition to behave	Anger is a disposition to attack the object of anger	J. B. Watson, B. F. Skinner, G. Ryle, N. Frijda
Cognitivist tradition	…a special way to appraise	Anger is the appraisal that a slight has been committed against me	Stoics, M. Arnold, R. Solomon, M. Nussbaum, R. Lazarus
Evolutionary tradition	…a special way of dealing with fundamental life tasks	Anger is the adaptive solution to the life task of fighting for survival	C. Darwin, S. Tomkins, R. Plutchik, P. Ekman, C. Izard, J. Tooby and L. Cosmides
Social constructionist tradition	…a special way of playing a social role	Anger is a social role in which one engages when wanting to be justified in the exercise of aggression	J.-P. Sartre, J. Averill, C. Lutz, R. Harre', P. Griffiths, B. Parkinson

Figure 1: Five traditions in the study of emotions

According to the *feeling tradition* (chapter 2), which finds its roots in common sense and is well exemplified by the likes of Aristotle, Rene Descartes, David Hume and William James, the emotions are essentially *ways of feeling*, i.e. special states of consciousness. For example, anger could be defined by a feeling theorist as the perception of a state of unpleasant arousal

characterized by trembling, increased heart beat, blood pressure, and breathing rate. The feeling theory was taken largely for granted in emotion theory roughly until the beginning of the 20th century. Updated versions of the feeling theory, which try to accommodate some of the criticisms launched against it in the course of the 20th century, have recently been proposed by Antonio Damasio (1994, 1999, 2003) in neurobiology and Jesse Prinz (2004a, 2004b) in philosophy.

According to the *behaviorist tradition* (chapter 3), developed at the beginning of the 20th century by the psychologist John Broadus Watson (1919, 1925) and further articulated by Burrhus Frederic Skinner (1953) in psychology and Gilbert Ryle (1949) in philosophy, the emotions are essentially dispositions to *behave*. A behaviorist may define anger as a disposition to attack or otherwise harm the object of one's anger. Even though purely behaviorist theories of the emotions collapsed together with behaviorism around the mid-1950s, traces of a behavioristic understanding of the emotions can be found in several contemporary theories. The psychologist Nico Frijda (1986), for example, defines the emotions as *action tendencies* of a particular sort. Frijda's account will be the main inspiration for my own theory of emotions, which I present in the constructive part of this work (chapter 10).

According to the *cognitivist tradition* (chapter 4), anticipated by the Stoics but articulated mostly in the 1960s and 1970s by philosophers such as Anthony Kenny (1963) and Errol Bedford (1957) and psychologists such as Magda Arnold (1960), Stanley Schachter and Jerome Singer (1962), the emotions are essentially ways of cognizing the world, commonly spelled out in terms of judgments or thoughts or appraisals. For example, under this tradition anger may be defined as the appraisal that one has been slighted. Updated versions of this approach, still very popular in emotion theory although under siege, have been offered among others by Robert Solomon (2003) and Martha Nussbaum (2001) in philosophy and Richard Lazarus (2001) and Klaus Scherer (2001) in psychology.

According to the *evolutionary tradition* (chapter 5), pioneered by Charles Darwin (1872) and brought to fruition in the 1960s by Silvan Tomkins (1962), Robert Plutchick (1980) and the affect program group gathered around Paul Ekman (1969, 1987, 1992, 1999b), the emotions are essentially mechanisms to deal efficiently with fundamental life tasks. For example, this tradition may identify anger with an adaptation to deal with recurrent conflict situations.

Finally, according to the *social constructionist tradition* (chapter 6), which emerged in the 1980s with anthropologists such as Catherine Lutz (1988), philosophers such as Rom Harre' (1986) and psychologists such as James Averill (1980, 1986), emotions are essentially culturally specific *social roles*. For example, a social constructivist may hold that anger is a social role taken on to be justified in the exercise of aggression. By being overcome by anger, people manage to get away with violating norms against aggression for the sake of norms that entitle them to the protection of their rights (Averill 1980, 66).

Updated versions of this theory have been offered by the psychologist Brian Parkinson (1995, 2005) and by the philosopher Paul Griffiths (2004a), who has integrated social constructionism with insights from the ethological literature. Their approach, which I label socio-evolutionary, combines insights from evolutionary and social constructionist traditions.

In chapter 7, I consider the two most popular theories in contemporary philosophy of emotions, namely Nussbaum's (2001) and Solomon's (1976, 2003) cognitivism and Prinz's (2004a, 2004b) and Damasio's (1994, 1999, 2003) neo-Jamesianism. I argue that they make the same mistake, namely trivialization by overextension. This mistake results from trying to fulfill at the same time two desiderata that are better kept apart, namely ordinary language compatibility and theoretical fruitfulness.

The first seven chapters of this dissertation make it apparent that every attempted definition of emotions offered within the five main traditions of research on emotions can be met by counterexamples. These counterexamples are commonly dealt with in one of four ways. Counterexamples consisting of purported cases of emotions which fail to meet the proposed definition are dealt with by arguing either that the definition, once properly interpreted, is in fact met, or that the purported case of emotion is in effect not an emotion. Counterexamples consisting of purported cases of non-emotions which meet the proposed definition are dealt with by arguing either that the definition, once properly interpreted, is in fact not met, or that the purported case of non-emotion is in effect an emotion. My diagnosis of these argumentative strategies is that they are entirely ad hoc, and stem from failure to understand the ground rules of the activity of theorizing about emotions.

The original sin of emotion theory, as I see it, is lack of methodological self-consciousness. Part 2 of my dissertation is an attempt to offer a methodology for emotion theory. I begin in chapter 8 with an investigation of the empirical literature on emotion concepts, namely on mental

representations of folk emotion categories. This literature reveals that folk emotion categories are highly heterogeneous and vague. This conclusion ought to be the starting point for theorizing about emotions, namely the realization that emotions as ordinarily understood comprise all kinds of items, and admit of borderline cases whose membership to folk emotion categories, or lack thereof, no amount of investigation will settle.

In chapter 9, I discuss what I take to be the desiderata for two central projects in emotion theory, namely the Folk Emotion Project, which aims to offer a *descriptive account* of categories such as "emotion" and "anger", and the Explicating Emotion Project, which aims to offer *explications* for them. I argue that a folk emotion theorist ought to aim for a *cluster account*, which makes explicit the properties such that fulfilling enough of them provides membership to the folk category. A cluster condition of membership strikes me as the best way to accommodate the sorts of empirical facts which led me to conclude (in chapter 8) that folk emotion categories are highly heterogeneous and have blurred edges. A theorist engaged in explication, on the other hand, ought to achieve similarity in use between his favorite explicatum and the folk explicandum, and in the process either reduce vagueness or increase fruitfulness or both. My account of the desiderata of explication follows to a large extent Carnap's (1950) original treatment, although with a couple of twists (chapter 9).

I am personally interested in the Explicating Emotion Project, because I am interested in scientific psychology, and I argue in chapter 9 that folk emotion categories are not natural kinds with respect to the explanatory and predictive practices of scientific psychology. This view has been prominently advocated by Griffiths (1997). I argue that Griffiths' adversaries have not fully understood what is implied by the claim that emotions are not natural kinds.

In chapter 10, I present the Urgency Management System theory of emotions, according to which an "umotion" is a special type of superordinate system for the management of situations of urgency endowed with pragmatic intentionality. I develop my theory of emotional intentionality from Millikan (2004), although the application to umotions I propose is new.

According to the UMS theory of emotions, emotions can be type-identified by a combination of an action tendency with control precedence and the conditions of appropriateness for the mechanism producing it. The history of selection of this mechanism is what establishes what the "umotion" represents, along the lines of teleosemantic theories of content.

In chapter 11, I offer a brief recap of what has been achieved, and a conclusion.

2. EMOTIONS AS FEELINGS

One of the most widely shared intuitions about the emotions is that they are passively experienced feelings, special states of consciousness by which emoters are overcome. I consider this intuition misleading, most importantly because it prevents us from appreciating the important agential dimensions of emotions around which I will ultimately construct my own theory. The influence of the feeling theory on the history of the subject cannot be overestimated. Practically everyone since Ancient Greece to the beginning of the 20th century was a feeling theorist.[2] 20th century emotion theory, in turn, can best be understood as a sequence of reactions to perceived shortcomings of the feeling theory. To understand contemporary debates, in which intuitions of passivity still loom large, we need to get a handle on the deep roots of such intuitions, and explain the grip they still have on us.

Studying the history of the feeling tradition also allows us to appreciate what problems the assimilation of emotions with feelings has encountered throughout the ages, and what arguments have been proposed to solve them. At the end of our exploration, we will be able to distinguish what should be retained of the feeling tradition from what we ought to get rid of.

Since the feeling theory has monopolized emotion theory for approximately 24 centuries, I can't offer anything more that a few highlights from this long and reputable history. The criterion of choice I employ is prominence. I take Aristotle, Descartes, Hume and James to be the four most influential thinkers within the feeling tradition. In the contemporary literature on emotions, they are commonly cited and discussed. More often than not, however, they appear under the guise of caricatures. This approach hides both the true shortcomings of their views and the insights worth gleaning from them.

[2] Arguably, the Stoics are an exception, although they gave an important role to feelings in their theory. See Sorabji (2000) for a masterful treatment of emotions in the Ancient World.

2.1. FEELINGS IN THE ANCIENT WORLD

2.1.1. Aristotle

In an influential paper, Cooper (1999, 407) states that the Aristotelian theory of the emotions is best understood as a "preliminary…investigation that clarifies the phenomena in question and prepares the way for a philosophically more ambitious overall theory", which "as far as we know" Aristotle never got around to provide. This is not to say that Aristotle does not have many interesting things to say about the emotions, scattered among his ethical treatises (Nichomachean Ethics, Eudemian Ethics), his book on the nature of the soul (On the Soul), his writings on poetry (Poetics), and most prominently his work on the art of public speaking (Rethoric).

The most influential of Aristotle's ideas about the emotions is undoubtedly the general characterization of the category of emotion as *passion* (pathe') in the Nichomachean Ethics. *Pathe'* are contrasted by Aristotle with *praxeis*, namely actions, in the form of a dichotomy which, to all intents and purposes, is still with us. This dichotomy is grounded on the idea that whereas actions are things we *do*, "in respect of the passions we are said to be moved".

As pointed out by Kosman (1980), the distinction between praxeis and pathe' in Aristotle is a "special instance or a more general structural duality, that of *poiein* and *pashein*, doing and being done" (105). The same duality appears in the *Categories*, where *doing* and *being done* are characterized as two of ten possible *modes of being*. The core idea at the heart of the notion of passion is therefore that passions are those things which happen to us, or that we are acted upon by, or that we undergo, or that we cannot voluntarily control.

"We feel anger and fear without choice", writes Aristotle, whereas "we are masters of our actions from the beginning right to the end" (1114b31-32). An important caveat is added, namely that although our passions are not expressions of choice, the dispositions to undergo them are. This is because such dispositions are associated with *character*, something that according to Aristotle can be voluntarily shaped in time by means of a process of habituation (*ethismos*, 1103a28).

The idea that the emotions are things we are acted upon by appears over and over in the history of emotion theory, and it is embedded in the very metaphors we still use to speak about

them. For example, "we 'fall' in love, are 'consumed' by envy, 'haunted' by guilt, 'paralyzed' by fear" (Averill 1980, 267). Many of the adjectives we deploy to refer to the emotions are "derived from participles" (Gordon 1987, 373) - frightened, surprised, joyed, irritated, upset –, another sign of how ingrained the idea of passivity is in our ordinary conceptualization of the emotions.[3]

To say that the emotions happen to us, however, is not yet to have provided a viable identity condition for them, because there are innumerable happenings which are not emotions. If I am hit by a meteorite, or if I fall into a hole, something happens to me, but it is not an emotion (even though an emotion may follow such happenings). For several centuries, the task of emotion theorists has been finding a way to distinguish passions from other happenings.

The core intuition of representatives of the feeling tradition is that the emotions are those happenings characterized by a special conscious experience or sensation or subjective quality or what-it-is-like aspect. Most commonly, this conscious experience has been labeled as *feeling*, a term which comes from the Middle English "felen", a derivation from the Indo-European root "pal-", from which the Latin "palpare", i.e. to touch. Two features of this conscious experience have been singled out as crucial for the instantiation of an emotion. The first is the *valence* of the experience, namely whether it is *pleasurable* or *painful*. The second is the *bodily character* of the experience, namely the fact that it is accompanied by *bodily sensations*. Consider the following passages:

> By passions I mean appetite, anger, fear, confidence, envy, joy, friendly feeling, hatred, longing, emulation, pity, and in general the feelings that are accompanied by pleasure or pain (*Nicomachean Ethics*, 2.5. 1105b19-24)

> The passions are all those feelings that so change men as to affect their judgments, and that are also attended by pain [lupe] or pleasure [hedone] (*Rhetoric* 2.1 1378a21-22)

In the first passage, Aristotle characterizes the passions as those feelings that are pleasurable or painful, whereas in the second passage he adds that such feelings are accompanied by dispositions to mental actions, i.e. judgments. Aristotle's interest in how passions influence judgments was related to his rationale for studying them. Although Aristotle discussed the

[3] In the history of emotion theory, the idea that emotions "happen" to emoters has often being followed by a corollary, namely that they are irrational. Starting with the Stoic account of the passions, emotions have been assimilated with disruptive happenings, which a wise man ought to get rid of. Notably, this was not Aristotle's view, since he believed that virtue demanded both doing the appropriate kinds of actions and undergoing the appropriate kinds of feelings (*Nicomachean Ethics*, 2.6. 1106b15-1106b30).

passions in several of his works, his most comprehensive account of individual passions was presented in the *Rhetoric*.

The practical objective of the Rhetoric was to help public speakers become more persuasive, especially in the context of political oratory and lawsuits. The ability to control his own and the audience's passions, Aristotle thought, will make the orator more effective. This is because the passions influence the way the persuasiveness of a speech will be judged. As he put it, "our judgments when we are pleased and friendly are not the same as when we are pained and hostile" (*Rethoric*, 1356a14-15). Aristotle discusses in some detail the following twelve passions: *anger, calmness, friendliness, hatred, fear, confidence in the face of danger, shame, kindness, pity, indignation, envy,* and *emulation*.

Aristotle aimed to provide a characterization of them that would be of help to the orator, who will be able to maximize his persuasiveness by learning (1) what kind of frame of mind is typical of people who experience a certain passion, (2) what kinds of people are such that a certain passion is generally experienced towards them, and (3) what kinds of circumstances characterize the experience of a passion. This will give the orator a way to exercise strategic control on the passions. Three further passions are given a more perfunctory treatment, namely *shadenfreude*, described in association with envy (1386b34-1387a3), *contempt*, described as the opposite of emulation (1388b22-28), and a "*feeling of satisfaction*" elicited by assisting to the distress of people who deserve to be distressed such as murderers (1386b25-33).

Even though Aristotle is interested in the impact of the passions on judgment, the fundamental feature that characterizes the passions is that they are attended by *pleasure* and *pain*. Cooper (1999) remarks that "lupe [pain] and hedone [pleasure] indicate…the character of the emotions as psychic disturbances in which we are set psychically in movement, made to experience some strong affect" (416). He suggests that pain and pleasure, used in a variety of different ways in the Aristotelian body of work, should not be understood in the Rhetoric as mild attitudes of liking or disliking. Rather, they should be understood as a form of *intense feeling*, sometimes associated with bodily symptoms (e.g. throbbing, gnawing, contracting, etc.).

In *On the Soul*, Aristotle explicitly states that "all the affections of soul involve a body-passion, gentleness, fear, pity, courage, joy, loving, and hating; in all these there is a concurrent affection of the body" (*On the Soul* 1.1 403a16-19). The feeling of *pain* characterizes for Aristotle the following passions: anger, fear, shame, pity, indignation, envy and emulation.

Pleasure is mentioned in the discussion of shadenfreude, and with respect to the "feeling of satisfaction" for someone else's deserved distress. However, not all passions end up being explicitly associated with either pain or pleasure. For example, Aristotle does not explicitly indicate in what sense friendliness, kindness, confidence in the face of danger, contempt and hatred involve either pleasure or pain. In some cases, the context clarifies that the lack of mention is a probably an oversight (e.g. confidence in the face of danger clearly presupposes pleasure in Aristotle's account), whereas in other cases the issue is open for interpretation (e.g. the cases of hatred and calmness).

What seems certain is that pleasure and pain, jointly with their bodily underpinnings, play a key role in the Aristotelian theory of the passions, especially with respect to passions which are also prototypical emotions (e.g. anger, fear, shame, pity, indignation, envy, shadenfreude, and contempt). It would be a mistake, however, to enlist Aristotle as a proponent of the idea that the passions are mere feelings of pleasure and pain. This is because Aristotle considers the passions to be an inseparable combination of *matter* and *form*. A thorough analysis of these two notions, and of the role they play in Aristotle's metaphysics, lies outside the scope of this dissertation (see e.g. Modrak 1983, Witt 1987). The basic idea is that the *form* of an entity is what *causes* its underlying *matter* to be what it is. The relevance of this distinction for our purposes is that Aristotle applies it to distinguish between two possible approaches to the definition of the emotions. Writes Aristotle:

> Hence a physicist would define an affection of soul differently from a dialectician; the latter would define e.g. anger as the appetite for returning pain for pain, or something like that, while the former would define it as a boiling of the blood or warm substance surrounding the heart. The latter assigns the material conditions, the former the form or formulable essence (*On the Soul*, 403a29-403b2)

When it comes to providing his own definition of anger, however, Aristotle combines an account of the material and formal aspects of anger with an account of what causes it, showing that he thinks "anger should be defined as a certain mode of movement of such and such a body (or part or faculty of a body)) by this or that cause and for this or that end" (403a25-28). In the Rhetoric, Aristotle writes:

> Anger may be defined as an impulse, accompanied by pain, to a conspicuous revenge for a conspicuous slight directed without

> justification towards what concerns oneself or towards what concerns one's friends (*Rhetoric* 2.2 1378a31-1378b1)

This definition illustrates that Aristotle's theory is ultimately a very rich hybrid that combines several of what in chapter 1 I called the *marks of emotionality*. Anger is defined as being *caused* by an *emotional appraisal* of "a conspicuous slight directed without justification towards what concerns oneself or towards what concerns one's friends" accompanied by *pain* (a conscious experience with a bodily underpinning, e.g. the boiling of a "warm substance surrounding the heart") and by a *behavioral disposition* (an impulse to a conspicuous revenge).

If we consider that Aristotle previously characterized the passions as "all those feelings that so change men as to affect their judgments", we realize that his definition of the passions also comprises *dispositions to mental actions* (e.g. judgments of persuasiveness).

Aristotle's definitions of the other twelve passions I listed before is as meticulous as his definition of anger, even though the account of anger is the only one in which Aristotle makes explicit what the relevant *behavioral dispositions* associated with the passion are (i.e. acting so as to get conspicuous revenge). This is instead left implicit – but clearly suggested - in the treatment of the other passions. For example, fear is defined as "pain or disturbance due to imagining some destructive or painful evil in the future" (1382a23), shame is defined as "pain or disturbance in regard to bad things, whether present, past or future, which seem likely to involve us in discredit" (1383b15), envy is defined as "pain excited by the prosperity of…people who are like us or equal with us."

I spent some time emphasizing the feeling component of the Aristotelian theory because Aristotle is often co-opted by contemporary cognitivists as their earliest honorary ancestor. The preliminary account I provided should be sufficient to show that, although Aristotle did give a prominent role to appraisals, he characterized the passions as being essentially feelings caused by appraisals and leading to actions.

Aristotle's least successful characterization seems to be that of *passion* as a superordinate category. When he speaks about passions in general, rather than of specific passions, Aristotle offers three characterizations: (a) they happen, (b) they are either pleasant or pleasurable, and (c) they tend to modify judgment. This account is not narrow enough, however, because there are many things which have these properties and are clearly not passions.

This is the case, for example, for sensory perceptions and bodily perceptions, which happen to us, can be either pleasant or painful and tend to influence judgment. Moreover, Aristotle's

account left it an open question why we should think that the passions are *feelings* of pleasure and pain in the first place.

Later representatives of the feeling theory tried to better characterize the class of passions within the class of sensations and to explain why a passion should be thought of as a sensation of a particular kind. I have chosen three authors to illustrate the developments of the feeling theory: Rene Descartes (1650), David Hume (1739), and William James (1884, 1890, 1894). Their theories well illustrate some of the difficulties of the feeling theory that ultimately led to its demise after the mid-20th century.

2.2. FEELINGS IN THE MODERN WORLD

Descartes (1650/1955) and Hume (1739/1992) both defined the passions as special kinds of conscious experiences, and tried to distinguish them from two other species of conscious experiences, perceptual experiences and bodily sensations. Their accounts are often singled out by critics as good examples of the incapacity of feeling theorists to understand that the emotions have intentionality. This criticism, as I will argue, is well-deserved, even though it should not blind us to the important insights offered by Descartes and Hume on the nature of the passions. I will show that, far from thinking of the passions as *mere* feelings, both authors were well aware of their multicomponential nature.

2.2.1. Descartes

In the *Passions of the Soul* (1650), Descartes tried to carve out the class of passions within the larger class of sensations, by arguing that the passions are the only conscious experiences we cannot be wrong about. Descartes' idea was that the passions, by not purporting to represent, cannot fail to represent. This idea finds its roots in the Cartesian distinction between three types of perceptions which are "found in the soul" and "caused by the body". According to Descartes, *body* and *soul* are two different substances, which interact in the pineal gland through the flow of

animal spirits. The body is an extended and non-thinking substance, whereas the soul is a non-extended and thinking substance.

Some of the perceptions found in the soul and caused by the body are referred to external objects ("nous les rapportons …aux objets de dehors"), some are referred to the body or some of its parts, and some are referred to the soul itself (art. 22).[4] The first are sensory perceptions (e.g. visual experiences, auditory experiences), the second are bodily sensations (e.g. pain, thirst, hunger), and the third are passions properly intended:

> The perceptions which are referred only to the soul are those whose effects are felt as if in the soul itself, and of which normally no proximate cause is known to which they can be attributed. Such are the sentiments of joy, anger, and others like them, which are sometimes excited in us by the objects which move our nerves, and sometimes also by other causes (art. 25)
>
> The passions of the soul are perceptions, sentiments, and emotions of the soul, which are referred particularly to the soul itself, and which are caused, entertained, and strengthened by some movement of the animal spirits (art. 27).

Descartes calls the passions "perceptions" or "sentiments" or "emotions of the soul", but he points out that the best way to designate them is the latter, in French "émotions de l'âme". Emotion comes from the Latin e-movere, which means to remove or displace. The French *émotion* is probably a derivation from *emouvoir*, a term in Middle French (14th-16th century French) which means "to stir up". Descartes' emphasis on favoring "emotion" with respect to other possible designations suggests that he ascribed to them the property of being, at least when intense, especially "disturbing". As Descartes put it, of all the perceptions a soul can have, "there are none that agitate it and disturb it so strongly as the passions" (art. 28).

This disturbing quality is primarily due to their bodily underpinnings, an interpretation Descartes suggests in various places. He claims for example that "the passions are nearly all accompanied by some disturbance which takes place in the heart and consequently throughout the blood and the animal spirits" (art. 46), a passage clearly hinting at physiological underpinnings associated with emotions (but notice the qualifier "nearly all").

The main ground of difference between passions and other sensations, however, is for Descartes a matter of *epistemic access*. He seems to think that when we perceive a bear, or a pain

[4] All Descartes' quotes are referred to the articles in which they can be found. There are 212 articles in *The Passions of the Soul*.

in the toe, we are implicitly making a causal hypothesis, respectively that there is an external bear, and that pain is located in the toe. In both cases, our causal hypotheses may be wrong, in the sense that there could be no bear in the external world, and the pain may not be located in the toe. In the most radical of cases, we could be dreaming up the entire experience.

But in the case of the passions, Descartes thinks, "normally no proximate cause is known to which they can be attributed" (art. 25), and therefore no fallible causal hypothesis is formulated. This passage is puzzling, because Descartes has told us that the animal spirits are the proximate cause of the passions. Could he be saying that what is normally not known is the *distal* cause of the passions? This would clearly be false, in the sense that on many occasions we know what the distal causes of our emotions are. Consider becoming afraid because you see a bear, or getting angry because someone stepped on your toe causing you pain. It seems natural to say in such cases that fear and anger are not referred "to the soul itself", but to the external bear and to whatever caused the pain in the toe.

But if that is the case, why do a visual sensation of a non-existent bear or a bodily sensation of a non-existent pain count as mistaken, whereas fear of a non-existent bear and anger about a non-existent pain do not? Descartes answer is ultimately that the passions are "so close and so interior to our soul that it is impossible that they should be felt without their being in reality just as they are felt" (art. 26). For example, "[e]ven if a man is asleep and dreaming, it is impossible that he should feel sad, or feel moved by any other passion, without it being strictly true that such passion is in the soul" (art. 26).

By denying that a dreamed passion and an actual passion differ in the same way in which a dreamed vision/bodily sensation differ from an actual vision/bodily sensation, Descartes is implicitly suggesting that the emotions do not purport to refer to anything they could fail to refer to. This amounts to denying that the distinction between actually representing and only purporting to represent applies to the emotions. This is a grave shortcoming for a theory of the emotions, and it is the main ground of the cognitivist critique of the feeling theory (see Deigh 1994).

As I will argue in chapter 4, the emotions have intentionality or the capacity to represent, and a good theory of the emotions ought to be able to account for it. However, it would be inappropriate to summarize the Cartesian position by claiming that Descartes understood the emotions as *mere* feelings. He did in effect define the general category of *passion* is a way that

goes dangerously close to saying that a passion is a mere feeling. However, when it came to defining specific passions, Descartes showed to be well aware of their multidimensionality. As he put it, to simply say that a passion is an agitation whose proximal cause is the movement of animal spirits "does not enable us to distinguish between the various passions: for that, we must investigate their origins and examine their first causes" (art. 51).

Descartes thinks that "objects which stimulate the senses" are the most common cause of the emotions, and proceeds to characterize the *appraisals* and *behavioural dispositions* of what he calls the six primitive passions: *admiration, love, hatred, desire, joy and sadness* (art. 69). According to Descartes, the primitive passions are the only passions which are neither subspecies of other passions nor obtainable from a combination of other passions. Descartes also offers accounts of non-primitive passions such as hope, jealousy, remorse, envy, anger, pride, disgust and others. Descartes attention for *behavioural dispositions* is related to his conviction that "the principal effect of all the human passions is that they move and dispose the soul to want the things for which they prepare the body" (40). Let us briefly consider three accounts of primitive passions.

Descartes tells us that "wonder" (art. 70) is a "sudden surprise of the soul" which is "primarily caused by the impression we have in the brain which represents the object as rare, and as consequently worthy of much consideration…" (appraisal) and which leads to "consider with attention the objects appearing rare and extraordinary" (mental disposition). Descartes remarks that "wonder" is a peculiar passion, in the sense that "we do not find it accompanied by any change in the heart or in the blood, such as it occurs in the case of the other passions" (art. 71). The idea here suggested is that wonder is the only passion lacking a detectable bodily underpinning.

Love and hatred (art. 79) are "emotions of the soul" – stirrings - caused by movements of the animal spirits brought about by evaluating an object as, respectively, agreeable and disagreeable. Love and hatred "incite" the emoter to, respectively, join with or separate from the object of the passion (behavioral disposition). Desire is "an agitation of the soul caused by the spirits which dispose it to wish for the future [behavioral disposition] the things which it represents as agreeable [appraisal]" (art. 86).

The structure of these definitions is similar to that of Aristotelian definitions. In both cases, pride of place is given to the *appraisals* which brings about the emotion (but Aristotle did a more

accurate job in describing them). Concerning differences, Aristotle in most of his definitions made the valence of the feeling (pleasurable or painful) explicit, whereas Descartes appeared to emphasize only the disturbing character of the feeling and its bodily underpinnings. Finally, Aristotle left implicit the behavioral tendencies associated with the passions (with the exception of anger), whereas Descartes gave them a central role, as he related them to the very function of the emotions.

2.2.2. Hume

Hume's theory, presented in *A Treatise on Human Nature* (1739/1992), offers a more sophisticated taxonomy of the *kinds of passions* than any theory formulated before it. Hume distinguished the passions from other sensations differently from Descartes, but he also concluded that they have no intentionality or representational purport.

According to Hume (1739), perceptions can be of two different kinds. On the one hand, there are *impressions*, which are those perceptions "which enter with most force and violence" into the mind. On the other hand, there are *ideas*, which are "the faint images of [impressions] in thinking and reasoning" (1). The passions are impressions, and so are sensory perceptions and bodily sensations.

What makes the passions different from other impressions, says Hume, is that whereas sensory perceptions and bodily sensations arise "in the soul originally, from unknown causes", the passions are "derived in a great measure from our ideas", which is why Hume calls them *impressions of reflection*. Hume introduces a distinction between two kinds of passions, the *calm* and the *violent* ones. This distinction is important, because it plants the seed for a theory of the emotions which does not require them to be conscious experiences.

Hume argues that "there are certain calm desires and tendencies, which, though they be real passions, produce little emotion in the mind, and are more known by their effects than by the immediate feeling or sensation" (471). In such class, he puts the "the sense of beauty and deformity" (276), "benevolence and resentment", "love of life", "kindness to children", or "the general appetite to good, and aversion to evil, considered as such" (417). Violent passions are instead exemplified by "the passions of love and hatred, grief and joy, pride and humility" (276). The calm passions, differently from the violent ones, produce little "emotion in the mind", namely little feeling-based disturbance. We can notice here that Hume is using the expression

"emotions in the mind" as Descartes used the expressions "émotions de l'âme", namely to designate an intense stirring. The calm passions are rather "known by their effects", which are for Hume primarily effects on behavior. If this is the case, then the calm passions could in principle be defined in terms of what they dispose us to do, whether or not any sensory experience is associated to them.

This is not the path chosen by Hume, who maintains that the calm passions are impressions, just impressions with a phenomenal quality which may at times be "almost imperceptible". This caveat reveals how deep the identification of the passions with feelings has run in the history of emotion theory. Hume, who laid the groundwork for speaking of the passions in non-phenomenological terms, never abandoned the idea that there is no passion without at least a modicum of sensation. To relinquish this idea demands calling into question a primitively compelling intuition neither Descartes nor Hume were prepared to question, namely that emotions are essentially kinds of *percepta*.

Hume's theory was similar to Descartes' theory is another crucial respect, namely the assumption that the passions lack the (non-derivative) ability to represent. This thesis was formulated by Hume (1739) along the following lines:

> A passion is an original existence, or, if you will, modification of existence, and contains not any representative quality, which renders it a copy of any other existence or modification. When I am angry, I am actually possest with the passion, and in that emotion have no more a reference to any other object, than when I am thirsty, or sick, or more than five foot high. It is impossible, therefore, that this passion can be opposed by, or be contradictory to truth and reason; since this contradiction consists in the disagreement of ideas, considered as copies, with those objects, which they represent (415).

This passage was offered in support of Hume's trademark thesis that *reason* and *passion* are not, and should not be, in conflict, in the sense that reason is, and should be, the slave of the passions. The slavery of reason concerns the direction of the will to act, which is for Hume the exclusive domain of the passions. According to Hume, "reason alone can never be a motive to any action of the will…[and]…it can never oppose passion in the direction of the will" (414). It cannot be a motive because according to Hume reason only delivers judgments that something is

the case (e.g. judgments that a certain effect will be caused by a certain action). These judgments cannot motivate unless supplemented by some passion which manages to turn, say, an expected effect into "the prospect of pain or pleasure".

Hume thinks that many have wrongly assumed that reason alone can motivate in part because some of the motivating passions are *calm*. Since such passions "cause no disorder in the soul, they are very readily taken for the determinations of reason, and are supposed to proceed from the same faculty, with that, which judges of truth and falsehood" (417). The idea is that the sensations produced by the calm passions are so inconspicuous that they can be confused with the operation of reason. This is another example of the fact that Hume thinks of the calm passions primarily as motivations to act, and puts their phenomenological aspect in the background.

The claim that the passions cannot represent is offered in support of the conclusion that reason cannot even oppose the passion "in the direction of the will", let alone direct the will unaided by passion. As Hume puts it, reason could only oppose a passion if a passion purported to represent something by being a copy of it. But the passions are impressions "derived in a great measure from our ideas", without being ideas themselves. When a passion is experienced, Hume remarks, it has "no more a reference to any other object, than when I am thirsty, or sick, or more than five foot high" (415). We can notice two differences with the Cartesian thesis that the emotions fail to refer.

The first is that Hume denies that bodily sensations such as thirst refer, whereas Descartes had claimed that they refer to the body. The second is that Hume is aware of the fact that the emotions can be warranted or unwarranted, namely that they have dimensions of normative assessment. In this respect, fear of a bear in a dream and fear of an actual bear would differ in the sense that the former would be unwarranted and the latter warranted. However, this does not prove that the emotions have "any representative quality", because what does the representing is for Hume always a judgment. According to Hume, there are only two cases in which "any affection can be called unreasonable" (416).

The first case is when "a passion, such as hope or fear, grief or joy, despair or security, is founded on the supposition or the existence of objects, which really do not exist" (416). Examples of this irrationality may be being afraid of a non-existent bear as in a dream, or grieving about the death of a non-existent relative.

The second case is when "in exerting any passion in action, we chuse means insufficient for the designed end, and deceive ourselves in our judgment of causes and effects" (416). Examples of this type of irrationality may be that of trying to hurt an enemy we hate by delivering him a gift he appreciates.

Hume states that if a passion is neither "founded on false suppositions", nor it brings about an action that fails to achieve the end, "the understanding can neither justify nor condemn it" (416). This being the case, "[i]t is not contrary to reason to prefer the destruction of the whole world to the scratching of my finger" (416). This represents some progress with respect to the Cartesian account, in the sense that, although derivatively, a passion can be unwarranted by Hume's own lights.

The problem is that the passion is, to paraphrase Hume, always a *slave of reason* when it comes to its normative assessment, because a passion as such does not have any non-derivative representational quality. When we speak of a passion as being unreasonable, we are always referring to the judgment accompanying it. The passion as such is "an original existence" which "contains not any representative quality". Hume is therefore as incapable as Descartes to realize that there may be something a passion aims to achieve *constitutively*, rather than by courtesy of a judgment that may be associated to it. This being said, it would be a mistake to accuse Hume of reducing passions to *mere feelings*. The complexity of the Humean account is best understood, once again, by considering the descriptions he offers of specific passions.

Hume distinguishes between *direct violent passions*, which "arise immediately from good or evil, from pain or pleasure" (e.g. desire, aversion, grief, joy, hope, fear, despair and security) and *indirect violent passions*, which arise "from the same principles, but by the conjunction of other qualities (e.g. pride, humility, ambition, vanity, love, hatred, envy, pity, malice, generosity) (399). He thinks all violent passions "both direct and indirect, are founded on pain and pleasure" (438), the removal of which brings about the removal of the passion. When pain and pleasure bring about a passion without mediation, the passion is direct. For example, "a suit of fine cloaths produces pleasure from their beauty; and this pleasure produces the direct passions, or the impressions of volition and desire" (439).

Volition is listed by Hume among the direct violent passions, but he points out that it is not to be considered a proper passion, a point I will disregard henceforth. *Desire* and *aversion* are the direct passion caused by respectively pleasure and pain "considered simply". I take "considered

simply" to mean that desire and aversion are what is elicited respectively by pleasure and pain when we do not take into account whether they are certain or uncertain. When pleasure and pain (or good and evil) are certain or very probable, the direct passions they cause are respectively *joy* and *grief*. When pleasure and pain (or good and evil) are instead improbable, the direct passions they cause are respectively *fear* and *hope*.

But suppose now that the beautiful "cloaths are considered as belonging to ourself" (439), namely that an agreeable *idea* of self is attached to them. In such case, the *indirect passion of* pride is generated. Hume characterizes the indirect passions in terms of their sensations (pleasurable or painful) and in terms of the ideas which, respectively, cause the passion and are caused by it. Writes Hume (1739):

> We must…make a distinction betwixt the cause and the object of [indirect] passions; betwixt that idea, which excites them, and that to which they direct their view, when excited (278).

Consider pride and humbleness, which are according to Hume "contrary" passions in that they allegedly cannot be experienced at the same time because one is pleasant and the other is unpleasant. Such passions have the same object: once they are generated, they bring about the idea of "self" in the emoter. From this Hume concludes that they must have different *causes*, because otherwise the same cause would generate both of them at the same time, and by assumption they cannot co-exist.

It is at this juncture that the important role of *appraisals* appears in the Humean theory of the passions. What causes pride rather than humbleness is the way a certain subject – e.g. cloaths - is *evaluated*. If the cloaths are evaluated as beautiful and belonging to oneself, pride ensues. If they are evaluated as ugly, and belonging to oneself, humbleness ensues. Every indirect violent passion is for Hume "derived from" a "double relation of ideas and impressions" (286).

The *relation of ideas* is that between the idea-cause and the idea-object of a passion. In pride, the idea-cause is a favorable idea of the self, and the idea-object is the self. In humbleness, the idea-cause is an unfavorable idea of the self and the idea-object is also the self. In love, the idea-cause is a favorable idea of another, and the idea-object is "some sensible being external to us". In hatred, the idea-cause is an unfavorable idea of another, and the idea-object is once again "some sensible being external to us" (329).

The *relation of impressions* is that between the sensation caused by the idea-cause – Hume says "independently" caused - and the way the passion as a whole feels. Hume considers the

"peculiar emotions [that idea-causes] excite in the soul" to be the "very being and essence" of the passions (286). In pride and love, the sensation is pleasant, whereas in humbleness and hatred the sensation is unpleasant.

The Humean account of the indirect passions, which I won't further analyze, is a description of the sensations, causes and objects of a variety of indirect violent passions. For example, he says of pride that "[a]ny thing, that gives a pleasant sensation, and is related to self, excites the passion of pride, which is also agreeable, and has self for its object" (288). This indicates that Hume's theory of the passions is also a hybrid, organized around the idea of a *double relation* of ideas and impressions.

The essence of the passions is said to be a conscious experience of pain and pleasure, but the Humean passions, as in Descartes' case, cannot be distinguished from one another merely by virtue of the sensations associated to them (e.g. love and pride do not differ in terms of their characteristic relation of impressions).

It is at this point that the relation of ideas - the idea-cause causing the passion (appraisal) and the idea-object caused by it - becomes critical for distinguishing the passions from one another. Love and pride are both pleasant, but the causal relations in which the pleasant sensation is embedded – being caused by a certain idea-cause and causing a certain idea-object - are different.

But what about the other marks of emotionality? Hume's theory is at its least successful when it comes to characterizing the role of *behavioral dispositions*. He claims that some passions lack such dispositions. Pride and humility, for example, are "pure emotions in the soul, unattended with any desire, and not immediately exciting us to action". On the other hand, other passions such as love and hatred are "not compleated within themselves, nor rest in that emotion, which they produce, but carry the mind to something farther" (367). As Hume puts it, "[l]ove is always followed by a desire of the happiness of the person beloved" (367) and hatred by a desire for his or her misery.

The idea that pride and humility are unattended by any desire is ungrounded. For example, pride will bring about the desire for being given credit for what one is proud of, for receiving praise, for sharing pride with other people, and innumerable other desires. Hume's aversion for considering behavioral dispositions and their underlying desires as part of the emotion goes deeper than just thinking that, as a matter of fact, some passions do not have any desires attached

to them. Even in the case of love and hatred, Hume is convinced that the desires by which they are always followed "are not the same with love and hatred, nor make any essential part of them" (368).

This shows that the Humean position is importantly different from the Cartesian one, which took the very function of the passions to lie in the way they prepared the body for action. Hume has two arguments – not clearly distinguished - for the conclusion that the desires associated with love and hatred, and more generally the passions, "are not absolutely essential to love and hatred".

On the one hand, they are not essential because, given the nature of our minds, we may love or hate "without our reflecting on the happiness or misery of their objects" (368). On the other hand, "[i]f nature had so pleased, love might have had the same effect as hatred, and hatred as love. I see no contradiction in supposing a desire of producing misery annexed to love, and of happiness to hatred" (368).

The problem with these arguments is that they presuppose that the *relation of ideas* required, by Hume's own lights, to identify the indirect passions is not constitutively related to the desires and behavioral dispositions associated with them. Hume assumes that desires and behavioral dispositions may be changed while maintaining unaltered the identity of the passion. If the identity of the passion changed as well, namely if the evaluations and sensations associated with it changed, then Hume would only be making the trivial claim that "love" could be the name given to what we currently call hatred and "hatred" could be the name given to what we currently call love.

I take Hume to be making a deeper point than that, namely that nature might have established that love is characterized by the idea-cause of some other person as agreeable, by the idea-object of some other person, by a pleasant sensation, and by the desire for the misery of the person appraised as agreeable. Under this view, a passion could be type-identified independently of the desires and behavioral dispositions associated with it. The problem with severing the relation between appraisal, emotion and associated desires and behavioral dispositions is that such relation is not arbitrary. A passion combining an idea-caused of agreeableness, an idea-object of some other person and a pleasant sensation could hardly be associated with a desire for the misery of the person appraised as agreeable. This sort of passion would lack psychological reality, and most importantly would not correspond to what we call "love".

It is indeed true that, as suggested by Hume, we can fail to "reflect on the happiness" of those we love, and possibly neglect them, but this proves only that love comes with a *tendency* to desire the happiness of the object of one's love, which may or may not be manifested. This is a much weaker claim than saying that love could be attached to no desire at all, or to any desire whatsoever and still be what we call "love", if nature had so pleased. As I will argue in chapter 10, there is no obstacle to assuming that behavioral tendencies are constitutively associated with emotions and at the same time that they can give rise to a wide range of behaviors or be inhibited.

2.3. PHYSIOLOGICAL FEELINGS

2.3.1. James and Lange

Descartes and Hume stated that the passions are percepta, and they pointed out that they generally amount to intense disturbances. Descartes and Aristotle clearly suggested that they have a bodily underpinning. Hume and Aristotle emphasized their valence, pointing out that all passions are pleasant or unpleasant sensations. Descartes and Hume admitted the possibility of some exceptions to their general accounts, but they considered them to be exceptions of degree rather than of kind. For example, Descartes pointed out that "wonder" may be the one passion without bodily underpinning, but he stuck to his account that every passion is a form of felt disturbance. Hume introduced an entire class of passions, the calm ones, which had almost imperceptible sensations associated to them, but he maintained that every passion is ultimately a felt impression.

The central question none of these authors had really addressed is: Why should a passion or emotion necessarily be a disturbing conscious experience? This question was explicitly answered for the first time by William James (1884, 1890, 1894) and Karl Lange (1885/1922), who independently offered the first theory of the emotions grounded in physiology. It is not an overstatement to say that 20th century emotion theory is a sequence of reactions to what is generally referred to as the James-Lange theory of the emotions, even though James and Lange

developed it independently. The theory has recently been revamped by a wave of influential Neo-Jamesians theories, defended by Antonio Damasio (1994, 1999, 2003) and Jesse Prinz (2004a, 2004b) among others, so it is worth paying close attention to it.

The inspiring thought of the James-Lange theory was that a truly scientific theory of the emotions required understanding them as essentially *physiological* rather than *psychic* phenomena. James argued that emotions had been described until then as "the internal shadings of emotional feeling", where feelings were understood as "psychic entities". Lange characterized the approach he meant to criticize as the "hypothesis of the psychical nature of the emotions", namely the view that an emotion is an event "of a purely psychical nature", or an "event in the soul", which in turn causes the bodily phenomena, which "are nevertheless in and of themselves wholly unessential". Both characterizations are broadly applicable to, among others, the theories of the passions offered by Aristotle, Descartes and Hume.

James (1890) believed that theorizing about the emotions in terms of internal shadings of emotional feeling resulted in endless classification, because "it is plain that the limit to their number would lie in the introspective vocabulary of the seeker, each race of men having found names for some shade of feeling which other races have left undiscriminated" (485). But this endless classifying was perceived by James as lacking in scientific rigor, because "you feel that [the] subdivisions [of descriptive psychology] are to a great extent either fictitious or unimportant, and that its pretences to accuracy are a sham" (448). What the scientific theory of emotion needed was for James a "central point of view, or a deductive or generative principle", which would never be discovered as long as the emotions were regarded as "absolutely individual things" hostage to the vagaries of introspective labeling (448). Since according to James "the general causes of the emotions are indubitably physiological", his conclusion was that by focusing on physiology he could find the "central point of view" he was looking for (448).

James compared the impact such principle could have on our understanding of emotion to the impact the "generative principle" of heredity and variation had on the understanding of biological species. Lange had a different complaint against the hypothesis of the psychical nature of the emotions, namely that such hypothesis is explanatory moot, namely not "indispensable for the explanation of the group of phenomena which we call emotions". From a scientific point of view, Lange argued, there is no need to think of the emotions as psychic entities. According to

both authors, what a scientific theory of emotions truly needed was to focus on the suite of *physiological responses* associated with the emotions, a *mark of emotionality* which had not until then to been put center stage in the definition of the emotions (although many authors had clearly hinted at it, e.g. Aristotle and Descartes).

I will illustrate the James-Lange theory by focusing on the Jamesian version, presented in "What is an Emotion?" (1884), and updated in the chapter on emotion of *The Principles of Psychology* (1890), and in "The Physical Basis of Emotion" (1894). I will mention Lange's contribution only when it integrates a point made by James, disregarding the minor differences between the two theories.

James distinguished two classes of emotions, the *standard* or *coarser* emotions on the one hand, and the *intellectual* or *subtler* emotions on the other. The former are those "in which every one recognizes a strong organic reverberation" (1890, 448), namely a "wave of bodily disturbance of some kind [which] accompanies the perception of the interesting sights or sounds, or the passage of the exciting train of ideas" (1884, 189). In this class, James included "surprise, curiosity, rapture, fear, anger, lust, greed", as well as "grief, .., rage, love". The class of subtler emotions is the class of "those [emotions] whose organic reverberation is less obvious and strong", and it includes "moral, intellectual, and aesthetic feelings", as well as "feelings of pleasure and displeasure, of interest and excitement" (1890, 448).

The distinction between standard and intellectual emotions is reminiscent of the Humean distinction between violent and calm passions, although Hume did not claim that the intense disturbance of the violent passions necessarily has a bodily underpinning. James began developing his "physiological theory" of the emotions on the basis of the *coarser emotions*, and then moved on to discuss the *subtler* ones. The natural way of thinking about the emotions, James pointed out, is that "the mental perception of some fact excites the mental affection called the emotion, and that this latter state of mind gives rise to the bodily expression" (1890, 449). James criticized this commonsensical approach on two main accounts.

Firstly, he argued that in most cases of emotion there is no mental affection between the mental perception of some fact and the bodily expression, in the sense that "*the bodily changes follow directly the* PERCEPTION *of the exciting fact*" (1890, 449, emphasis in original). Secondly, he argued that the emotion is precisely "*our feeling of the same changes as they occur"* (1890, 449, emphasis in original). This amounts to a reversal of common sense,

according to which "we lose our fortune, are sorry and weep; we meet a bear, are frightened and run; we are insulted by a rival, are angry and strike" (1890, 449-450). According to James, instead, "we feel sorry because we cry, angry because we strike, afraid because we tremble" (James 1894, 189-190).

James' theory demanded supporting evidence of two kinds. On the one hand, evidence that bodily changes can follow the perception of the exciting fact *directly*. James and Lange believed that evidence to this effect could be interpreted as an existence proof that the emotions are not *essentially* mental affections causing bodily changes. Secondly, it required an argument to the effect that an emotion is nothing but the perception of bodily changes.

Most of the critical literature on the James-Lange theory has focused on attacking the thesis that an emotion is nothing but the perception of a bodily change. Generally, debunking this thesis has been considered sufficient to debunk the thesis that "*the bodily changes follow directly the* PERCEPTION *of the exciting fact*". But when we read the evidence James provided for the latter thesis, we realize that it *does not* presuppose the truth of the thesis that emotions are perceptions of bodily changes.

The thesis James in effect defended ought to be reformulated as the thesis that "emotions *follow directly the* PERCEPTION *of the exciting fact*". It is with respect to this thesis that James made what I consider some of his most important and least discussed contributions.

In a nutshell, James offered a number of reasons why we should not assimilate *emotional appraisal* to the forms of *intellectual evaluation* of which a deliberate judgment that something is the case is a paradigmatic example. Call this the *Argument from Direct Elicitation* against the traditional way of conceiving of emotions. As I will argue in chapter 7, James foreshadowed some of the central criticisms launched against the cognitivist theory of emotions 70 years later. The Jamesian account is not systematic, but full of important leads. For example, James noticed that emotional evaluations are often very *fast*: "[i]f we abruptly see a dark moving form in the woods, our heart stops beating, and we catch our breath instantly and before any articulate idea of danger can arise" (1890, 457).

He also noticed that they are sometimes *insulated from rational considerations*: "[i]f our friend goes near to the edge of a precipice, we get the well-known feeling of "all-overishness," and we shrink back, although we positively *know* him to be safe" (1890, 457). In other cases, emotions do not even seem to *result from* evaluations of *specific objects*: "The best proof that the

immediate cause of emotion is a physical effect on the nerves is furnished by *those pathological cases in which the emotion is objectless*" (1890, 458).

Lange mentioned two further cases to "prove the superfluous nature" of the hypothesis that emotions are psychic entities, namely that of emotions induced in infants by loud noises and emotions induced by chemical substances (e.g. alcohol, bromide). What the heterogeneous examples offered by James and Lange suggest is that the experience of emotion – i.e. the experience of bodily changes as far as they are concerned - is brought about at least sometimes *directly* by the "perception of the exciting fact".

The perception of the exciting fact amounts in effect to a process of appraisal which does not require time to be executed, is not penetrable by and/or available to rational cognitive processes, and is available to infants and, we may add, to animals.

As James put it, certain emotions suggest that "peculiarly conformed pieces of the world's furniture will fatally call forth most particular mental and bodily reactions, in advance of, and often in direct opposition to, the verdict of our deliberate reason concerning them" (1894, 191).

James and Lange did not offer a theory of the intentionality of the emotions, even though - differently from Descartes and Hume - they neither explicitly nor implicitly denied that emotions have representative qualities.

By suggesting a number of important distinctions between emotional and non-emotional appraisals, they pointed our attention to some of the fundamental phenomena an account of the intentionality of the emotions should be able to accommodate (e.g. their fastness, their availability to creatures without language, etc.). Jesse Prinz's (2004a, 2004b) recent Neo-Jamesian theory of the emotions, which I discuss in section 7.1, can be seen as an attempt to endow the Jamesian theory with intentionality while preserving its trademark thesis, namely that an emotion is a perception of bodily changes.

Let us now turn to an analysis of how James defended such key thesis. The following passage, one of the most often cited in the history of emotion theory, illustrates what James took to be his main source of evidence. I call this the *Argument from Conceivability*:

> I now proceed to urge the vital point of my whole theory, which is this. If we fancy some strong emotion, and then try to abstract from our consciousness of it all the feelings of its characteristic bodily symptoms, we find we have nothing left behind, no "mind-stuff" out of which the emotion can be constituted, and that a cold and neutral state of intellectual perception is all that remains…*What*

> *kind of an emotion of fear would be left, if the feelings neither of quickened heart-beats nor of shallow breathing, neither of trembling lips nor of weakened limbs, neither of goose-flesh nor of visceral stirrings, were present, it is quite impossible to think…* A purely disembodied human emotion is a nonentity…[F]or *us*, emotion dissociated from all bodily feeling is inconceivable (1884, 194, emphasis in original)

This is an very influential argument, so let us focus on it in some detail. A virtue of James' argument is that it sheds light on what James takes a *bodily change* to be for the purposes of the thesis that an emotion is nothing but the perception of a bodily change. It is obvious that the truth of the thesis depends on what one means by a "bodily change". In his first formulation of the thesis, James stated that "we feel sorry because we cry, angry because we strike, afraid because we tremble", and called all such things "bodily manifestations".

Under this view, an *expression* such as crying, an *instrumental behavior* such as running, and a *physiological response* such as trembling would all count as *bodily manifestations* the perception of which is the emotion. Before discussing whether or not this is a good interpretation of what James meant by bodily manifestation, let us notice that James, often presented as holding the view that emotions are *just* perceptions of bodily changes, was well aware of some of their other marks of emotionality, e.g. *expressions* (e.g. crying) and *behavioral dispositions* (e.g. running). James was also aware of the presence of emotional *appraisals*, which he described as direct (Argument from Direct Elicitation).

I am convinced that the best interpretation of what James meant by "bodily changes" is not the all-encompassing one I have just sketched. An overly liberal notion of bodily changes seems to me in conflict with the starting point of the James-Lange theory, which was to focus on emotions as *physiological* phenomena. James explicitly described the theory he and Lange endorsed as a "physiological theory" (1890, 449), a qualifier suggesting that the bodily changes whose perception is the emotion are to be interpreted as *physiological bodily changes*. James did not ask: What would be left of fear if we didn't run, or if we didn't display the facial expression of fear. Rather, he asked what would be left of fear without "quickened heart-beats", "shallow breathing", "trembling lips", "weakened limbs", "goose-flesh", "visceral stirrings", and so on. These are all *physiological responses of the autonomic variety*, namely changes governed by the autonomic nervous system. This suggests that "the vital point of my whole theory" is for James

that an emotion is a perception of autonomic changes, rather than a perception of expressions or instrumental behaviors.

In conclusion, I want to argue that the common turn of phrase with which the James-Lange theory is generally summarized, namely "we are sad because we cry", must be considered ambiguous between two interpretations. On the one hand, the thesis that "we are sad because we cry" states is that we are sad *insofar* as we cry, in the sense that "we" cannot conceive of sadness without a perception of sadness-typical bodily changes (Argument from Conceivability). On the other hand, the thesis that "we are sad because we cry" states that - at least sometimes - we are sad without there being a specific thought causing our sadness (Argument from Direct Elicitation). Such thought may be absent when the emotion is objectless, or too quick to allow the formation of a thought.[5] On occasion, sadness may even occur in opposition to a thought, as when we get sad about things we believe should make us happy.

2.4. CONCLUSION

In this chapter, I discussed Aristotle's pioneering account of the passions and what I consider to be the three most important emotion theorists respectively of the 17th (Descartes), 18th (Hume), and 19th century (James). I have undertaken an analysis of prominent feeling theorists for two reasons. On the one hand, I wanted to react to what I consider to be widespread misunderstandings concerning their theories. My main point was that it is patently false that feeling theorists thought of the passions as *mere feelings*. Surprisingly, this is the common way in which their position is portrayed in contemporary emotion theory (see chapter 7). All the feeling theorists I discussed understood that emotions are commonly caused by appraisals and comprise behaviors and behavioral dispositions. On the other hand, as I will argue in chapter 4, traditional feeling theorists lacked insight into the idea that there is a form of aboutness constitutively associated with emotions. In a sense we will have to understand, emotions are endowed with representational qualities, precisely the sorts of qualities Descartes and Hume denied them.

[5] This may not work too well for sadness, but it certainly does for fear

The other reason why I focused on the feeling theory is that it is the vantage point from which the history of 20th century emotion theory needs to be understood. The main traditions of research of the last 100 years have all defined themselves in contrast to the feeling theory. The feeling theory has been criticized for its reliance on the method of introspection hailed by Descartes as infallible (behaviorism, chapter 3), for its neglect of the cognitive dimension of emotion (cognitivism, chapter 4), for its incapacity to account for the commonality of emotions across species (evolutionary tradition, chapter 5), and for its blindness to the social and communicative dimensions of emotional phenomena (social constructionism, chapter 6). My task in the next four chapters will be to reconstruct how these competing traditions came to light, what intuitions and what arguments propelled their emergence, and what substantive problems ultimately emerged to limit their influence.

3. EMOTIONS AS BEHAVIORS

Behaviorism brought about a radical change in the method of psychology, and produced an approach to the emotions which lasted roughly from the beginning of the 20th century to the late 1950s. In this chapter, I do not mean to provide a general characterization of behaviorism, nor of its profound impact on a variety of scientific disciplines, nor of the reasons underlying its ultimate demise (see Boakes 1984 for a history of behaviorism). I only aim to highlight the general structure of the approach championed by behaviorists with respect to the emotions.

Because of their intuitive connection with *feelings*, the emotions constituted one of the main challenges behaviorists had to face. When Watson (1925) began addressing the topic, James was singled out as the main representative of what a theory of the emotions should never be. The designation of James as the root of all evils in emotion theory, as we shall see, is a common rhetorical tool in the 20th century, and it testifies both to the profound impact of his theory, and to the unscrupulousness of some of his enemies. My attempt to clarify what James *really* said was meant precisely to expose his many caricatures, some of which feature in the behaviorist literature as well.

I follow the customary distinction between two different strands of behaviorism, *psychological behaviorism* and *philosophical behaviorism*. Psychological behaviorism, championed most prominently by Watson (1925) and Skinner (1953), had a fundamentally scientific motivation. It was argued by that, *for the purposes of science*, what counts as an emotion can only be an externally observable manifestation, whereas internal states of consciousness – e.g. feelings – ought to be expunged from the subject matter of the psychology of emotions.

Philosophical behaviorism, championed most prominently by Gilbert Ryle (1949), had instead an eminently philosophical motivation, namely the analysis of *the way we talk* about the emotions. The importance of Ryle's behaviorism is that it pointed us to a feature which had

previously not been fully understood, namely that there are very many different uses of the term "emotion" in ordinary language. Ryle argued that most of them refer to dispositions rather than occurrent feelings. This claim became the centerpiece of a general attack against a Cartesian view of the mind as the "ghost in the machine" of the body.

The historical investigation of behaviorism is particularly interesting for my project, because the theory of emotions I propose in chapter 10 considers the impact emotions have on behavior to be their essential feature. Understanding the behaviorist articulation of this idea will allow me to become aware of some of its possible uses, as well as some of the mistakes to be avoided. Also, I consider the Rylean exploration of the heterogeneity of the ways in which we talk about emotions to have important methodological consequences, namely that it is very hard to offer a single theory of emotions that fits them all. To make this point persuasively, however, we need to expand on Ryle's introspective study of emotion talk with empirical evidence on the way in which people use emotion categories, a task I leave for chapter 8.

3.1. PSYCHOLOGICAL BEHAVIORISM

3.1.1. Watson and Skinner

There is a remarkable similarity between the concerns voiced by James and Lange against their predecessors and those voiced by the psychological behaviorists against James and Lange. As the reader will recall, James and Lange had rejected as unscientific the identification of emotions with psychic states. According to James (1890, 449), "the trouble with the emotions in psychology is that they are regarded too much as absolutely individual things", endlessly classifiable in terms of introspectable shades of feeling "without ever getting on to another logical level…[w]hereas the beauty of all truly scientific work is to get to ever deeper levels" (1890, 449).

The way out from the "level of individual description" was to think of them in terms of perceptions of bodily changes, and focus on bodily changes as the "central point" of view need for a scientific psychology of emotions. Lange added that, for the purposes of a scientific

psychology, there was no need to think of the emotions other than in terms of their physiological underpinnings. Watson and Skinner submitted the notion of *perception of bodily change* to the same criticism to which James and Lange had submitted the notion of *psychic state*. They said that it was not scientific, and that it was not needed for the purposes of a truly scientific psychology. The contentious notion was not that of a bodily change as such, but that of its *conscious perception*. Watson echoed James's own concerns about the unprincipled multiplication of individual analyses of inward feelings of emotion with respect to the notion of consciousness itself. He wrote in "Psychology as the Behaviorist Views it" (1913), a.k.a. the behaviorist manifesto:

> As a result of this major assumption that there is such a thing as consciousness and that we can analyze it by introspection, we find as many analyses as there are individual psychologists. There is no way of attacking and solving psychological problems and standardizing methods (5).

Watson's conclusion was that a truly scientific psychology should get rid of the notion of consciousness entirely. This demanded a revolutionary change in the nature of psychology, understood by James as the science which aims for "the description and explanation of states of consciousness as such" (6). According to Watson, this kind of psychology "has failed to make good its claim as a natural science" (1913, 9), and ought to be reformed with respect to it *subject matter*, *goal*, and *methodology*. The subject matter of psychology ought to be "facts of behavior" (9), and its "theoretical goal is the prediction and control of behavior" (1). This being the case, its method cannot be introspection of conscious states. Rather, "psychology…is a purely objective, experimental branch of natural science which needs introspection as little as do the sciences of chemistry and physics" (9).

In 1913, he presented his distrust of consciousness as grounded in the impossibility of its scientific study. When he came back to the topic of consciousness in *Behaviorism* (1925), however, Watson had significantly radicalized his position. Consciousness had moved from the status of *entity to be ignored* for the purposes of science to the status of *non-entity*, as he wrote that "belief in the existence of consciousness goes back to the ancient days of superstition and magic" (1925, 2). The focus on consciousness had, according to Watson, made the James-Lange theory of the emotions unscientific.

With characteristic curtness, Watson wrote that "nearly 40 years ago James gave the psychology of the emotions a setback from which it has only recently begun to recover" (1925,

140). The problem is that "each man has to make his own introspections. No experimental method of approach is possible. No verification of observation is possible. In other words, no scientific objective study of emotion is possible" (1925, 142). Skinner (1953) echoed such concerns, and pointed to another problem for states of consciousness, namely their causal mootness. As he put it, introspective psychology "defines its 'subjective' events in ways which strip them of any usefulness in a causal analysis. The events appealed to in early mentalistic explanations of behavior have remained beyond the reach of observation…" (1953, 30-31).

"The emotions", Skinner (1953, 160) argued, "are excellent examples of the fictional causes to which we commonly attribute behavior". Focusing on internal variables unavailable to scientific investigation has led psychologists to neglect "[t]he external variables of which behavior is a function". The solution was for both Watson and Skinner to focus on the causal relationship between the dependent variable of *behavior*, that which psychology has the theoretical goal of controlling and predicting, and the independent variables constituted by *external conditions*.

In effect, psychological behaviorists recommended the external stimulus-behavioral response pattern as a general blueprint for the study of what had previously been thought to be facts of consciousness as such. With respect to the emotions, this led to trying to understand them in terms of stimulus-response sequences. This strategy, however, presented an immediate difficulty, namely that the same emotion at different times and in different people can result from many different stimuli and it can be manifested by many different behaviors. How can we capture this multiplicity within the stimulus-response schema?

Watson's response came in two stages. Firstly, he assumed that in infants, the multiplicity of emotion-eliciting stimuli is dramatically reduced, and so is the multiplicity of unconditioned emotional responses to such stimuli. Secondly, he assumed that the variety of emotions present in adults resulted from a process of conditioning on the basis of the unconditioned emotional responses present at birth. Since he took the goal of psychology to be the control and prediction of behavior, he assumed that the task of a psychologist of the emotions was to master the control of emotional responses, and shape them in socially advantageous ways.

This meant revealing what are the emotional stimuli which generate unconditional emotional responses, how they lead through process of conditioning to the complexity of adult emotional life, and how the process of conditioning can be interfered with at will. Watson

distinguished three unconditioned stimulus-emotional response sequences present at birth, which he called *fear*, *rage* and *love*. The following chart shows which unlearned responses were called by Watson (1925) fear, rage and love, and which stimuli were supposed to produce them (he added that there might be other producing stimuli, but that in any case they were "few in number"):

Emotions	Emotional stimuli	Unconditioned emotional responses
Fear	Loud sounds, loss of support	"Checking of breathing, jump or start of the whole body, crying, defecation and urination (and many other not worked out experimentally. Probably the largest group of part reactions are visceral)" (1925, 156)
Rage	Restraint of body	"Stiffening of the whole body, screaming, temporary cessation of breathing, redding of face changing to blueness of face, etc. it is obvious that while there are general overt responses, the greatest concentration of movement is in the visceral field" (1925, 156)
Love	Striking skin and sex organs, rocking, riding on foot	"Cessation of crying, gurgling, cooing and many others not determined. That visceral factors predominate is shown by changes in circulation and in respiration, erection of penis, etc." (1925, 157)

Figure 2: Watson's theory of fear, rage and love

Under this view, infant fear/rage/love are instantiated by responding to the observable stimuli in the second column with the observable "behaviors" in the third column. Allegedly, every adult emotion results from infant fear, rage and love through a process of conditioning. Under this view, fear, rage and love are primitive emotions in the following sense: every other emotion is obtained from them by conditioning. An adult emotion, say adult guilt, is a conditioned response obtained by successive conditioning of fear, rage and love responses to

novel stimuli. This is a very unconvincing theory, both because some adult instances of love, fear and shame do not seem to fit the conditioning paradigm (e.g. falling in love may take years) and because it is entirely mysterious how emotions such as, say, shame, guilt and awe could result from either fear, rage or love or any combination of them by classical conditioning. I will disregard these problems, focusing only on the idea that an emotion can be defined by stimulus-response sequence.

We should notice that Watson's list of emotional responses in infants does not comprise conscious experiences (e.g. pain, bodily sensations), including instead physiological responses (e.g. visceral responses, defecation and urination, stiffening) and expressive/instrumental behaviors (e.g. crying, screaming). If we subtract the perception part, and substitute the appraisal of the stimulus with the stimulus itself, we are not too far from James's own theory (under a broad interpretation of bodily change). Watson's theory of emotions is ultimately that an emotion is a particular kind of bodily manifestation (understood broadly enough to comprise physiological and behavioral responses) to particular stimuli.

There are two main reasons why behaviorists ended up focusing on behaviors rather than physiological responses. The first is that, at least with the instruments of the time, physiological events were not as accurately observable as were instrumental behaviors and expressions. In this respect, physiological events failed to meet the bar of scientific observability, not in principle as it was the case for conscious experiences, but in practice. The second reason is that it was widely believed that physiological changes were not sufficiently different from one another to distinguish the emotions from one another.

Skinner (1953) made the point explicitly, when he said that "[i]n spite of extensive research it has not been possible to show that each emotion is distinguished by a particular pattern of responses of glands and smooth muscles. Although there are a few characteristic patterns of such responses, the differences between emotions are often not great and do not follow the usual distinctions" (160-1).

This passage refers to evidence presented in 1929 by Walter Cannon in *Bodily Changes in Pain, Hunger, Fear and Rage*, a widely influential textbook in physiology. Cannon argued that the physiological differences between different emotions, at least those that could be measured at the time, were not sufficient to distinguish them from one another. I will come back to this influential thesis later on. Skinner (1953) had the same worry about facial and postural

expressions of emotions, as he said that "it has not been possible to specify given sets of expressive responses as characteristic of particular emotion" (161). His recommended solution was to focus on instrumental behaviors, which he assumed to be both observable without difficulties and such as to differentiate the emotions from one another.

Skinner's theory is that emotions are *dispositions* to respond to certain stimuli with certain behaviors. For example, he wrote that "When the man in the street says that someone is afraid or angry or in love, he is generally talking about predispositions to act in certain ways. The angry man shows an increased probability of striking insulting, or otherwise influencing injury and a lover probability of aiding favoring comforting or making love…so defined, an emotion…is not to be identified with physiological or psychic conditions" (1953, 163).

This remark is insightful, and it was fully developed by Ryle (1949), who argued convincingly that in many cases in which we ordinarily speak about the emotions, what we refer to are dispositions. However, it is certainly incorrect that "[t]he names of the so-called emotions serve to classify behaviors with respect to various circumstances which affect its probability" (1953, 162). The names of the emotions are often used to refer to occurrent events (e.g. behavioral events), rather than a disposition to engage in them with a certain probability.

The problem with the behaviorist theory of the emotions, however, is more general, and it is the problem that led behaviorism to its demise in the 1950s. Already in the 1920s and 1930s psychologists such as Edward C. Tolman and Wolfgang Köhler had argued against the stimulus-response paradigm for its dismissal of intervening variables, pointing out that the mentalistic notion of *purpose* was constitutive of the notion of behavior. In an influential review of Skinner's *Verbal Behavior*, Chomsky (1959) argued that the behaviorist definition of "verbal behavior" cannot be clearly characterized without reference to mental mechanisms generating it, more specifically mental grammars consisting of rules. Chomsky's critique can be generalized, in the sense that it seems very hard to characterize behaviors as being of the same kind without making implicit presuppositions about the mental processes that cause them and are caused by them.

This problem is present in the case of emotions. The distinction between, say, guilt and shame cannot be drawn exclusively in terms of stimulus-response sequences, in the sense that the very same sequences – e.g. greeting a friend with an apology - may in principle belong to both guilt and shame. One may argue that it is because they are species of the same kind, but

the similarities can take place even with respect to emotions as diverse as love and hatred. For example, one may run away from someone because of both unrequited love and hatred.

In this sense, Skinner was wrong in thinking that instrumental behaviors do not incur into the same problem in which physiological responses and expressions can incur, namely that they do not seem sufficient in themselves to type-identify the emotions. If we want to use behaviors to do so, as I will argue in chapter 10, we need to describe them in terms of their constitutive *purposes*, a notion which has no role in a strictly behaviorist account of the emotions.

Moreover, a given stimulus-response sequence does not even reveal whether or not an emotion is instantiated, let alone which emotion is instantiated. Most stimulus-response sequences which can feature in emotional episodes can feature in non-emotional ones as well. Finally, what appears to be missing from the behaviorist theory is an account of the fact that emotional behaviors often have an urgency lacked by non-emotional ones. In Skinner's theory, this feature is completely absent, as emotional behaviors are described as dispositions, without any consideration for the urgency by which emotional dispositions claim, at least sometimes, to be manifested. These mistakes will be avoided in the theory of emotions I will formulate in chapter 10.

3.2. PHILOSOPHICAL BEHAVIORISM

3.2.1. Ryle

Emotion terms are often deployed to designate dispositions. This is the central point about the emotions made by Gilbert Ryle in *The Concept of Mind* (1949), a book in which he attacked Cartesian dualism by arguing that the language of mental states, far from referring to a mysterious realm of private mental episodes, referred to *facts of behavior*. The Rylean project amounted to tackling various categories of mental states - the will, emotion, self-knowledge, sensation, and imagination, etc. – and show that they correspond to facts about how people do behave or would behave in certain circumstances.

The reason why I distinguish Ryle's philosophical behaviorism about the emotions from Watson and Skinner's psychological behaviorism (despite the obvious similarities especially with Skinner) is that the two projects had very different motivations.

For the psychological behaviorists, the central problem was that, since feelings are not intersubjectively observable, we should not offer a definition of emotions in terms of feelings. For Ryle (1949), the central problem was that, since feelings are not what ordinary speakers commonly refer to when they use emotion terms, we should not offer a definition of emotions in terms them. Writes Ryle (1949):

> There are two quite different senses of 'emotion', in which we explain people's behaviour by reference to emotions. In the first sense we are referring to the motives or inclinations from which more or less intelligent actions are done. In the second sense we are referring to moods, including the agitations or perturbations of which some aimless movements are signs. In neither of these senses are we asserting or implying that the overt behaviour is the effect of a felt turbulence in the agent's stream of consciousness. In a third sense of "emotion", pangs and twinges are feelings or emotions, but they are not, save per accidens, things by reference to which we explain behaviour (110).

All in all, Ryle distinguishes between four different entities the term *emotion* refers to in the ordinary language: (a) *feelings*, (b) *inclinations* (or motives), (c) *moods* (or frames of mind) and (d) *agitations* (or commotions). *Feelings*, he says, are "the sorts of things which people often describe as thrills, twinges, pangs, throbs, wrenches, itches, prickings, chills, glows, loads, qualms, hankerings, curdlings, sinkings, tensions, gnawings and shocks" (82). Ryle further characterizes feelings as "things that come and go or wax and wane in a few seconds; they stab or they grumble; we feel them all over us or else in a particular part" (97).

Inclinations are "motives by which people's higher-level behaviour is explained", for example the motives of "vanity, kindliness, avarice, patriotism and laziness" (82).

Moods are things such as being "depressed, happy, uncommunicative or restless" (95), which can last minutes, days or even a life-time (in the latter case, we will generally speak of character traits).

Agitations, finally, are things such as being "anxious, startled, shocked, excited, convulsed, flabbergasted, in suspense, flurried, and irritated" (90), which interfere with our thinking or acting, and which have feelings as their "sign".

Now, in his chapter on the emotions, Ryle argued for two theses: (a) Emotion terms often do not refer to feelings, i.e. turbulences in the stream of consciousness, (b) When emotion terms refer to feelings, such mental occurrences should not be construed as introspectable causes of purposive action. I will disregard the second thesis, which is not relevant for my purposes, and focus instead on the first.

Thesis (a) is the thesis that emotion terms are often best understood as referring to inclinations, agitations and moods, which are for Ryle (1949, 81) "not occurrences and do not therefore take place either publicly or privately. They are propensities, not acts or states". Ryle's strategy is to take as self-evident that inclinations, agitations, and moods are very often what we refer to when we speak about emotions, and argue that such things neither are nor presuppose "feelings". I will only consider Ryle's argument about inclinations.

To say of a man that he is vain, Ryle argues, is to say that he will "behave in certain ways", such as talking about himself, rejecting criticism, seek the footlights and so on. It is also to say that he will show certain patterns of thinking, such as indulging "in roseate daydreams about his own successes", or avoiding "recalling past failures and to plan for his own advancement" (83). Ryle acknowledges that, among the things a vane man is disposed to do, there may also be the experience of feelings. For example, "to have an acute sinking feeling, when an eminent person forgets his name, and to feel buoyant of heart and light of toe on hearing of the misfortunes of his rivals" (84).

But he wants to resist the idea that inclinations must be understood as referring *essentially* to dispositions to feel. For example, we may explain the actions of a certain man by claiming that he has an inclination towards Symbolic Logic. But such motive cannot be assumed to involve a disposition to "experience impulses of a peculiar kind, namely feelings of interest in Symbolic Logic" (86). Ryle plausibly remarks that "there are no peculiar feelings of interest in Symbolic Logic for him to report", namely no particular perturbation by which he is overcome when acting on the motive constituted by his interest for Symbolic Logic. From this Ryle concludes that "to do something from a motive is compatible with being free from any particular feelings while doing it" (85).

Moreover, Ryle remarks that having a certain motive may sometimes be psychologically incompatible with having a feeling associated with it. For example, if a vain man were to really

feel vane (as opposed to having other vanity-induced feelings other than that of vanity), his vanity would be undermined as a motive.

Ryle concludes that "there is no special thrill or pang which we call a 'feeling of vanity'", and adds that the vane man is likely to be the "last person to recognize how vain he was" (85). There is a problem in using examples such as these to make points about the nature of emotions, namely that it is unclear whether, under an ordinary understanding, either interest for Symbolic Logic or vanity count as emotions.

But I take Ryle's point to be ultimately a good one, which can be reformulated with respect to things that are unquestionably emotions. For example, when we explain the actions of a certain man by claiming that he feels guilty towards his mother, we are not implying that he has a disposition to experience feelings of a particular sort. Our ascription may be referring exclusively to a disposition to behave and think in certain ways (e.g. make amends, think of himself as a bad person, try to change, etc.).

Ryle makes a further important point that is relevant to discussions about the role played by feeling in emotion. The point is that the term "feeling" itself does not necessarily refer to bodily feeling. Ryle gives the following examples: "I feel ill", "I feel stupid", "I feel capable of climbing a tree", "I felt a lump in the mattress", "I felt cold", "I felt my chin with my thumb", "I felt in vain for the lever", "I felt as if something important was about to happen", "I felt that there was a flaw somewhere in the argument", "I felt that he was angry" (101-103).

In all such cases, it seems quite clear that we are not referring to the occurrence of a bodily feeling. As a matter of fact, we can paraphrase such expressions with roughly equivalent ones that do not mention feelings at all. For example, we can say "I believe I am ill", "I think I am capable of climbing a tree", "I touched a lump in the mattress", "I am cold", "I became convinced that there was a flaw somewhere in the argument", and so on.

Ryle suggests that there is something in common between uses of *feel talk* that are and that are not reports of feeling. In both cases, what is felt has not been "established by careful witnessing…or inferred from clues", but rather grasped without having done a specific investigation and accumulated evidence. As Ryle puts it, that "he felt it is enough to settled some debates; that he merely felt it is enough to show that [some] debates should not even begin" (103). Ryle concludes that the fact that we say things such as "I feel guilty" (an inclination) or "I feel depressed" (a mood) does not imply that what is referred to is a feeling,

in the same sense that when we say “I feel I can do it” does not imply that what is referred to is a feeling.

To say that often our emotion and feeling language does not refer to actual bodily feelings, however, is not to say that it never does. Ryle himself admits that sometimes we do refer to feelings, and he notices that the way in which we talk about emotions as feelings is very similar to the way in which we talk of bodily sensations. For example, we speak of twinges of remorse, but also of twinges of rheumatism, we speak of qualms of apprehension, but also qualms of sea-sickness, we speak of glows of pride but also glows of warmth.

Ryle suggests that both in the case of both feelings and in the case of bodily sensations, reporting them is making causal hypotheses. When someone speaks of experiencing a “twinge of toothache”, Ryle says, he is making a causal hypothesis, which may be wrong. For example, “[a] wounded soldier may say that he feels a twinge of rheumatism in his right leg, when he has no right leg, and when ‘rheumatism’ is the wrong diagnosis of the pain he feels” (101). Similarly, when someone reports a “twinge of remorse”, he is giving “a diagnosis of it, but a diagnosis which is not in terms of a physiological disturbance. In some cases his diagnosis may be erroneous; he may diagnose as a twinge of remorse what is really a twinge of fear”, or ascribe to “dyspepsia a feeling which is really a sign of anxiety” (102).

Ryle admits that “such mis-diagnoses are more common in children than in grown-ups, and in persons in untried situations than in persons living their charted lives”, but maintains his conclusion that “whether we are attaching a sensation to a physiological condition or attaching a feeling to an emotional condition, we are applying a causal hypothesis” (102).

This portion of Ryle’s argument is a direct response to the Cartesian attempt to distinguish the class of passions from the class of sensory and bodily sensations by claiming that we cannot be mistaken about the passions whereas we can about other forms of perception. This part of Ryle’s argument is specifically meant to attack the standard Cartesian picture of privileged introspective access, namely that idea that our passions “are so close and so interior to our soul that it is impossible that they should be felt without being in reality just as they are felt” (Descartes 1650, art. 17-26).

Ryle makes an interesting further point with respect to feelings. He says that “[n]either my twinges nor my winces, neither my squirming feelings nor my bodily squirmings, neither my feelings of relief nor my sighs of relief, are things which I do for a reason; nor, in consequence,

are they things which I can be said to do cleverly or stupidly, successfully or unsuccessfully, carefully or carelessly" (102). This theme, as we shall see in the next chapter, will become one of the trademark arguments against the feeling theory that emotions are mere feelings. The worry is that feelings are not the sorts of things to which normative properties can be ascribed. But if this is true, then emotions cannot be (mere) feelings.

3.3. CONCLUSION

In this chapter, I described the radical behaviorist shift that occurred in emotion theory at the beginning of the 20th century. I have distinguished between a psychological and a philosophical wing of behaviorism about the emotions.

The psychological wing was motivated by methodological considerations on the requirements a science of the emotions ought to fulfill. The philosophical wing was motivated by a desire to reject the Cartesian picture of emotions as feelings and the idea of infallible and privileged access to them. Psychological behaviorism made us acutely aware of the fact that introspection is an inductive practice, and that such practice can go wrong. But the specific behaviorist recipe of eradicating consciousness entirely from the subject matter of psychology was unduly restrictive. It resulted in about half a century of purgatory for the notions of cognition, feeling and purposive behavior, all of which are required to acquire a deep understanding of what the emotions are.

Ryle's account had the great merit of bringing to our attention the fact that there are very many things we mean when we speak about the emotions in ordinary language. In particular, Ryle made a strong case for the thesis that we often do not refer to occurrences at all, let alone feelings, but rather to disposition to behave.

4. EMOTIONS AS COGNITIONS

The cognitivist tradition finds its roots in Aristotelian and Stoic insights about the emotions, but it reached a mature stage only in the latter part of the 20th century. After the demise of behaviorism, the cognitivist theory of emotions was proposed as a solution to problems which, allegedly, neither behaviorism nor the feeling theory could solve. In the past forty years, it has become the dominant tradition in both philosophy and psychology, even though its central commitments are increasingly hard to stand by.

This chapter is devoted to reconstructing the main strands of the early cognitivist critique of the feeling theory, in particular the arguments offered to reject the assimilation of emotions with sensations that type-identify them. There is much to learn from such arguments, but I do not think they support cognitivism as a research program. What they do is to present us with desiderata that any good theory of emotions ought to fulfill.

As I see it, the cognitivist tradition fosters at least two serious misunderstandings, namely that emotions are necessarily cognitively sophisticated phenomena, and that their motivational dimension is secondary. The limits of the cognitivist tradition are paradigmatically expressed by the trademark thesis that emotions are judgments, a thesis Robert Solomon (2003) and Martha Nussbaum (2001) have energetically defended for the past twenty years. My discussion of contemporary cognitivism, however, will have to wait until chapter 7, because I want to offer a critique of it which also applies – mutatis mutandis - to cognitivism's main competitor, namely the increasingly popular Neo-Jamesian theory of emotions (Damasio 1994, 2003; Prinz 2004a, 2004b). My focus will be instead on the emergence of cognitivism as a research program.

Despite its shortcomings, cognitivism has left us with at least two noteworthy legacies. The first is the realization that emotions have constitutive conditions of appropriateness. For example, fear is the sort of thing which is inappropriate in the absence of danger. Naïve forms of behaviorism and of the feeling theory fail to explain this important fact about the emotions, and

they are consequently hopeless. The second legacy is the realization that it is not events in themselves that generate emotions, but rather the way they are appraised or evaluated or interpreted by emoters. This means that any theory of emotions must be able to account for the relation between the way events are "cognized" and the emotions that follow them.

In this chapter, I distinguish three classes of arguments offered against the feeling theory. The *Argument from Absent Consciousness* state that emotions cannot be feelings, because we often speak of emotions which lack consciousness, in various senses in which such notion can be understood. The *Argument from Intentionality* states that the emotions cannot be feelings because they have intentionality, a property theorists in the feeling tradition had either denied (e.g. Descartes, Hume) or neglected to account for (e.g. James). The *Argument from Differentiation*, finally, states that emotions cannot be distinguished from one another in terms of their feelings, because such feelings are not sufficiently differentiated.

4.1. THE ARGUMENT FROM ABSENT CONSCIOUSNESS

The Argument from Absent Consciousness states that emotions cannot be feelings, because emotions often lack concomitant conscious feelings. The argument comes in several varieties depending on how we understand the notion of consciousness.

4.1.1. Two Notions of Consciousness

The notion of consciousness can be understood in several different ways. A useful distinction is the one drawn by Block (1995) between *phenomenal consciousness* (P-consciousness) and *access consciousness* (A-consciousness). Writes Block:

> A state is P-conscious if it has experiential properties. The totality of the experiential properties of a state are "what it is like" to have it.
>
> A state is access-conscious if it is poised for direct control of thought and action (Block 1995, 380-382)

Whether or not a state has *phenomenal consciousness* depends upon whether or not there is a way it is like to undergo it, namely upon whether or not a subjective experience of a particular

kind is associated with being in that state. The paradigmatic example of a P-conscious state is pain, but any state whatever may in principle have a way it is like to undergo it (e.g. sensory perceptions, thoughts, etc.). Whether or not a state has *access consciousness* depends instead upon whether or not the state can be used by the agent for theoretical and practical inferences by simple redirection of attention (Block says that is must be "freely available"). Block thinks that the verbal reportability of a state is "often the best practical guide to A-consciousness", but he is prepared to grant A-consciousness also to the states of an animal, provided they rationally control thoughts and actions.

The paradigmatic example of A-conscious states in a human being are "states with representational content expressed by "that" clauses", for example the thought that one does not like fish, which can be reported, can be used in reasoning, and can influence action through a practical inference. There are several issues of interpretation concerning what exactly makes a state P-conscious or A-conscious, but I won't discuss them here, relying on a rough and ready understanding of Block's distinction.

Block's key suggestion is that, although states generally are either both P-conscious and A-conscious or neither, some states may be endowed with one type of consciousness but not the other. For example, consider a blindsight patient, who lacks both P-consciousness and A-consciousness with respect of the visual states in which he is in when an object is presented in his blind field. If we hypothesize that the blindsight patient becomes what Block (1995) calls a superblindsight patient, namely one who lacks a way-it-is-like to experience objects in his blind field, but who is poised to form thoughts such as "there is an object X in my blindfield", we have an example of a state which is A-conscious without being P-conscious. On the other hand, if we hypothesize that the blindsight patient is miraculously given a way-it-is-like to experience objects in the blind hemi-field, but not the ability to report on what he sees or otherwise use his visual experiences to think and act, we have a case of P-consciousness without A-consciousness.

The emotion theorists we have studied so far did not distinguish between access P-consciousness and A-consciousness of emotion. It is quite clear from the work of Aristotle, Descartes, Hume and James that they assumed the emotions to be both P-conscious and A-conscious. The P-consciousness of emotions was accounted for primarily in two ways, namely as being *valenced* and as being *bodily*. In other words, the way it is like to undergo an emotion was assumed to be a pleasurable or painful experience amounting to the perception of bodily

changes. The A-consciousness of emotion was simply taken for granted, in the sense that it was assumed that if there is a way it is like to undergo an emotion E, then one can always freely report "I am undergoing emotion E", and use this thought in reasoning and action.

The cognitivists questioned the assumption that the emotions are always P-conscious and A-conscious, pointing to the possibility of emotions endowed with just one kind of consciousness or lacking both. The natural starting point for thinking about emotions without consciousness is Freud, who first offered a systematic alternative to the view that consciousness is the mark of the mental.

4.1.2. Emotions without access-consciousness

4.1.2.1. The Freudian unconscious

As Guzeldere (1997) pointed out, "until the time of Freud, there was no proper theoretical framework in which the reject the Cartesian idea of equating the mind with whatever lay within the scope of one's consciousness" (Block et al., 1997). At its highest level of abstraction, the Freudian insight is that the mind is not exhausted by what mindful creatures are conscious of. As Freud put it, consciousness just represents the tip of an iceberg, whereas the bulk of mental processes occur below the surface of consciousness.

Freud (1915) did not explicitly define the consciousness of a state, but understood it roughly in terms of the joint presence of phenomenological consciousness and access consciousness. His most distinctive contribution was the distinction between two levels below the surface of consciousness, that of the *pre-conscious* and that of the *unconscious*. The *preconscious* includes all those states that are "capable of entering consciousness…without any special resistance and given certain conditions" (1915, 106), for example those which can enter consciousness by simple redirection of attention.

The *unconscious* includes instead all those states that have been repressed as unacceptable to the conscious mind. Unconscious states can become conscious by virtue of a process which eliminates repression, namely the psychoanalytic process. Later in his career, Freud (1923, 1940) added to his topographical account of the mind a division into three functional components: The *Id*, which contains all instincts present at birth (e.g. eros and thanatos), the *Ego*, which mediates

between Id and external world with the task of self-preservation, and the *Superego*, which limits the pursuit of pleasure in which the Ego is engaged in light of moral constraints apprehended during childhood. Even though the mapping is not perfect, the general idea is that the influence of Ego and Superego on behavior is both conscious and preconscious, whereas the influence of the Id is unconscious. The result of repressed conflicts between the three functional components of the mind is according to Freud the experience of anxiety.

Freud thought that *ideas* were the paradigmatic example of states that can be conscious, preconscious or unconscious, whereas his position on the possibility of unconscious *emotions* was nuanced. Freud thought that one should not speak of unconscious emotions, in the sense that "there are no unconscious affects in the sense in which there are unconscious ideas" (1915, 111). The reason for this claim is that Freud endorsed the view that "it is surely of the essence of an emotion that we should feel it, i.e. that it should enter consciousness" (1915, 110). Under this view, which well exemplifies the grip of the feeling tradition on Freud, to speak of unconscious emotion is to speak inappropriately.

But Freud acknowledged that "in psychoanalytic practice we are accustomed to speak of unconscious love, hate, anger, etc., and find it impossible to avoid even the strange conjunction 'unconscious consciousness of guilt', or a paradoxical unconscious anxiety'" (110). What is meant by unconscious emotion, Freud argued, is that certain emotions are "perceived, but misconstrued" (100), in the sense that "[b]y the repression of its proper presentation it is forced to become connected with another idea, and is now interpreted by consciousness as the expression of this other idea" (110).

I find it useful to employ Block's (1995) distinction in order to capture the Freudian position on the emotions. What I take Freud to be saying is that an emotion is always a phenomenological experience of a certain kind (by Freud's definition of emotion), but that sometimes emotions lack access consciousness, in the sense that while someone is experiencing, say, jealousy, he or she is not poised to form the thought "I am jealous", and use such thought in his mental activity and action. I think this is an important insight by Freud, but we need to liberate it from the idea that repression is the only or even the primary mechanism preventing an emotion from being access conscious.

Consider the following example by Bedford (1957), one of the first cognitivists to attack the view that "an emotion is a feeling, or at least an experience of a special type which involves a

feeling" (77). Bedford pointed out that "we can be mistaken about our own emotions, and that in this matter a man is not the final court of appeal in his own case; those who are jealous are often the last, instead of the first, to recognize that they are" (81-82).

We can notice here that, although Bedford addressed his example against the view that emotions are experiences, the jealousy example is best understood as an example of lack of access-consciousness. It is compatible with jealousy being felt, just not *as* jealousy. For example, the jealous man may experience butterflies in the stomach and an irregular breathing pattern during a party in which his girlfriend talks intently to another man, but ascribe his bodily sensations to the high room temperature or to tiredness. In such case, there would be a way it is like to undergo jealousy, but the emoter would lack access to the thought "I am jealous".

To explain this case by invoking the mechanism of repression disregards the fact that access to the thought "I am jealous" may be available just a few minutes after the jealousy episode, without any need for psychoanalysis. On the other hand, simple redirection of attention during the jealousy episode would not have been enough to form the thought "I am jealous now" by Bedford's assumption. Moreover, the idea that emotions which lack access consciousness always do so because of their conflict with instincts related to the libido and death has been widely criticized for being unsupported by hard evidence (e.g. Grunbaum 1984).

4.1.2.2. The cognitive unconscious

The abandonment of the idea that repression determines the boundary between the conscious and the unconscious seems to me the fundamental aspect of the shift between the *Freudian unconscious* and what Kihlstrom (1987) labeled the *cognitive unconscious*. When theoretical interest in the unconscious re-emerged in the 1950s, roughly around the time behaviorism collapsed, the unconscious started being broadly characterized as that portion of the mind which is not available for verbal report, yet exerts a demonstrable influence on some aspect of cognitive performance. Whereas Freud focused on unconscious *thoughts*, the focus of cognitive scientists was on *unconscious processing of information* in the course of a variety of cognitive performances (e.g. perception, memory, learning, thinking, speaking, etc.).

More generally, the focus shifted from the state of being in a certain mental condition unconsciously to the process of arriving unconsciously at a certain mental condition. Since a

large part of unconscious processing has nothing to do with repression, but is simply an effect of cognitive architecture, the cognitive unconscious is a much more wide-ranging notion than the Freudian unconscious, and it certainly cannot all be eliminated through psychoanalysis.

This way of thinking about the unconscious finds an early example in Hermann von Helmholtz's (1860/1979) claim that "[t]he psychic activities that lead us to infer that there in front of us at a certain place there is a certain object of a certain character, are generally not conscious activities, but unconscious ones". What is meant here is not that we cannot verbally report on what we perceive, but rather that we cannot verbally report on the mechanisms by which we perceive.

The notions of *unconscious processing* and *unconscious state* can of course combine, but even when they do so the lack of consciousness of a state need not have something do with repression. For example, if a target stimulus and a masking stimulus are presented in rapid succession (less than 30ms) in a backwards masking experiment (Marcel 1983), the experimental subjects report that they are not aware of being exposed to the masked visual stimulus (e.g. the picture of a snake).

In such case, they are not only unconscious of what Helmholtz would have called the inferences in which their perceptual system is engaged, but also unconscious of what they are perceiving, in the sense that they do not have access consciousness with respect to their percepta (and presumably not phenomenal consciousness either).

Often, the notion of *unconscious processing* is equated by cognitive scientists with that of *automatic processing* (e.g. Posner 1978), a view which finds its root in James' claim that a state can be unconscious in the sense that we do not have our attention directed on it (Freud would have called such state preconscious rather than unconscious).

As James (1890) put it, "my experience is what I agree to attend to" (402), the corollary of which is that what do not attend to is not part of my conscious experience. Shiffrin & Schneider (1977) distinguished between automatic and controlled information processing, pointing to differences between these two forms of processing in terms of the attentional resources, cognitive resources, effort, intentional control, and memory structures they mobilize. Ever since the distinction was introduced, the presence of automatic or unconscious information processing has been detected in a large number of cognitive activities.

Starting with Zajonc (1980), the presence of unconscious processing has been investigated also with respect to so-called "hot cognition", namely emotions, preferences and other affective states. This has ultimately led Kihlstrom and his associates to claim that "it is time for the new science of emotion to entertain the possibility of an emotional unconscious" (Kihlstrom, Mulvaney, Tobias, & Tobis, 1998).

The notion of the *emotional unconscious*, however, is dangerously ambiguous between two main interpretations. What is meant by some who use the expression "unconscious emotion" is that the nature of the information processing which brings about the emotion is not conscious (e.g. Zajonc 2000, Öhman 1999). This is the same sense in which we could say of visual perception that it is unconscious because, say, it relies on 2½D sketches (Marr 1982) or that speech perception is unconscious because it relies on Chomskian mental grammars. What is meant by others who speak about "unconscious emotion" is instead that the emoter is in a certain emotional state – however he got there – unconsciously (e.g. Berridge et al. 2003).

The two notions of "unconscious emotion" are orthogonal to one another. Whether or not the processing of information leading to an emotion is unconscious does not determine whether or not one will unconsciously be in a certain state of emotion (and viceversa). To eliminate ambiguity, I will reserve the expression *unconscious emotion* for emotions lacking access consciousness, namely emotions such that those who undergo them are not poised to formulate the thought that they are undergoing them by simple redirection of attention. I will instead speak of *unconscious emotional processing* when I want to refer to the special features of the etiology of an emotion (e.g. its automaticity).

Once we admit the possibility that there are emotions with respect to which emoters have no access consciousness, it becomes clear that Freud's hypothesis that an emotion essentially has P-consciousness comes into question. A remark in this spirit can be found in the work of the cognitivist Martha Nussbaum (2001), who writes that "[i]f we are prepared to recognize nonconscious emotional states, such as nonconscious fear of death or nonconscious anger…then we cannot possibly hold to any necessary phenomenological condition for that emotion-type" (61).

Unconscious fear of failing, to pick a classic example, is likely to be both A-unconscious and P-unconscious, in the sense that if the rationale for ascribing it is to explain a certain pattern of behavior inclusive of slips of the tongue, dreams, anxiety states, and so on, it becomes

mysterious why we should maintain the assumption that there must necessarily be a way it is like to undergo the consciously inaccessible process that generates such behaviors.

A different example of emotions lacking both access-consciousness and phenomenal consciousness is that of emotions understood as dispositions. As I argued in chapter 3, the idea that emotions are not essentially feelings because emotion terms sometimes refer to dispositions was convincingly defended by Ryle (1949).

Echoes of the Rylean critique can be found in the work of early philosophical cognitivists in the 1960s (e.g. Bedford 1957, Pitcher 1965). Pitcher (1965) pointed out for example that when we call someone jealous, what we are saying may be simply that he *would* do, think or feel in certain ways if the circumstances *were* such and such. A man at a party may be jealous in this dispositional sense even if his girlfriend is sitting alone in a corner, provided that he would respond with, say, butterflies in the stomach or an aggressive behavior or thoughts of punishment if another man were to talk charmingly to her.

4.1.3. Emotions without bodily phenomenology

As I pointed out, feeling theorists generally assumed that the way the emotions feel had a bodily nature. James defined an emotion as being the perception of a bodily change, a view hinted at by many in the history of emotion theory (e.g. Aristotle and Descartes). The cognitivists cast doubt on this idea, pointing out that sometimes either the emotions lack phenomenology all together or fail to have a specifically bodily phenomenology. We have already explored two examples of complete lack of phenomenology in the previous section, when we talked about unconscious emotions and emotions as dispositions, both of which lack both P-consciousness and A-consciousness (but we have considered the possibility that an unconscious emotion may have P-consciousness as presupposed by Freud).

In this section, I want to discuss examples of emotions that have access-consciousness but lack phenomenological consciousness of a bodily type. In cases such as “hope and envy”, says for example Nussbaum (2001, 3), “we can’t even begin to specify such a defining feeling”. Pitcher (1965) tells us that bodily sensations are “characteristic features of emotion situations-although only for some emotions, not for all”, and that in any event they are “not absolutely essential ones, so that there may be occasional emotion-situations which lack them” (339).

Solomon (2003) states that "autonomic nervous system responses are not an essential part of every emotion" (221).

The general thesis here is that most if not all emotions can be had in principle without concomitant bodily changes (e.g. anger), and that many emotions are very rarely if ever had with concomitant bodily changes (e.g. shame). As noticed by many in the history of emotion theory, there seems to be a large class of emotions which are commonly ascribed in the absence of bodily changes. Hume called them *calm passions*, and James called them *subtler emotions*.

Paradigmatic examples include moral emotions such as guilt, shame and disgust, intellectual emotions such as cerebral rapture in an intellectual topic, and aesthetic emotions such as delight at the sound of music. One of the tasks for theorists who believe we are capable of conceiving of emotions in the absence of bodily changes is to explain what would be left of an emotion once the bodily changes are subtracted.

In other words, one must answer James' challenge in the Argument from Conceivability, and explain what would be left of fear once we remove its autonomic underpinnings. The general answer is: every other mark of emotionality with the exclusion of the bodily one, namely an emotional appraisal, a conscious experience without bodily underpinnings, and a suite of expressions, instrumental behaviors and mental behaviors. Early cognitivists did not make much headway in explaining how emotions can be instantiated by virtue of these marks when the bodily changes are absent. They noticed what appeared to them as a self-evident fact, namely that we do often speak of emotions even in the absence of bodily changes. But we can find some interesting suggestions concerning the way in which emotions could be instantiated without bodily phenomenology.

The cognitivist Bedford (1957), for example, made some insightful comments, focusing his attention on what we do when we *ascribe* emotions to ourselves and others through *speech acts*. The advantage of thinking of emotional behavior in terms of speech acts is that by applying the taxonomy of speech acts elaborated by Austin (1962) and Searle (1969, 1980) to this particular kind of emotional behavior we can have a preliminary insight into what could be the point of ascribing an emotion if it is not necessarily that of reporting on a bodily change. Bedford (1957) argued that when one says "I feel shamed now" or "I am quite disgusted with the literary men", it's not the case that "the primary function of these statements is to communicate psychological

facts" (93). Rather, Bedford thinks that one is, respectively, making "an admission of responsibility, or perhaps a plea in mitigation" and "condemning literary men".

Let us explore this insight focusing on the sentence "I feel ashamed now". If we apply Searle's (1969, 1980) taxonomy of speech acts to it, we realize that there may be various illocutionary points to such utterance, namely various things the utterer may be trying to achieve by saying it. The speech act is assertive, as the utterer is stating that he appraises his behavior to have been blameworthy in the way shameful behaviors are (Bedford speaks of an "admission of responsibility"). The speech act is directive as well, in the sense that, as Bedford astutely notices, the utterer may be making "a plea in mitigation", trying to get the hearer to forgive him. The speech act, finally, is commissive, in the sense that it seems reasonable to assume that the utterer is expressing an intention to avoid repeating the behaviors he is ashamed of in the future. Can the utterer be occurrently ashamed without experiencing bodily changes? If he can sincerely and correctly ascribe shame to himself in ordinary English without experiencing bodily changes, the answer would have to be positive. And I do not see any obstacle to performing successfully any of the speech acts I described in the absence of bodily changes.

As it turns out, this is not a special case due to the fact that the manifestation of the emotion is a speech act. I will argue in chapter 10 that understanding what the emotions are requires understanding that they convey information about antecedent circumstances, expectations and intentions. It seems to me unquestionably true that this information can on occasion be conveyed without bodily underpinnings. The key to the emotional, as I will argue, is urgency, which is often but not always associated with bodily underpinnings.

4.2. THE ARGUMENT FROM INTENTIONALITY

Philosophical cognitivists further complained that feeling theories had failed to account for the *intentionality* of emotions. Broad (1954) is the first self-described cognitivist I am aware of to discuss in some detail the sense in which the emotions have intentionality (but already Brentano had included love and hatred in his list of paradigmatic intentional states). Broad distinguished between two kinds of experiences, those which do and those which do not have an

"epistemological object". Experiences of the first kind are such that there is something they are *about* or *of* or *directed towards*, whereas experiences of the second kind are such that undergoing them is not being "*aware of a certain object*, real or fictitious", but rather "feeling in a certain way" (Broad 1954, 283; emphasis in original).

Perceiving or *thinking* are paradigmatic experiences with an epistemological object, in the sense that perceiving and thinking are, at first blush at least, perceiving *of* something or thinking *about* something. On the contrary, *feeling hot* or *feeling tired* were singled out by Broad as being paradigmatic experiences without an object, in the sense that one is not hot or tired *about* things (although one may be hot or tired *because of* things). Emotions, Broad claimed, belong to the class of experiences with an epistemic object, since we are afraid *of* things, angry *at* people, guilty *about* our actions.

The argument proposed by Broad was that since feelings are experiences without epistemic objects, and emotions are experiences with epistemic objects, emotions cannot be feelings. This argument, however, is problematic. Its main limitation is that is presupposes that experiences either are about epistemic objects or they are feelings. A feeling theorist may quickly evade the argument by saying that the emotions are those feelings which do have epistemic objects. It is hard to imagine that Descartes, Hume or James would have denied that one can be afraid of things, or angry about things.

4.2.1. Kenny on formal objects

A more promising formulation of the Argument from Intentionality can be found in the work of Anthony Kenny (1963). His main critique to the feeling theory is summarized in the following passage:

> Descartes and Hume, with the philosophers and psychologists who followed them, treated the relationship between an emotion and its formal object, which is a logical one, as if it were a contingent matter of fact. If the emotions were internal impressions…there would be no logical restrictions on the type of object which each emotion could have. . It would be a mere matter of fact that people were not angered by being benefited, nor afraid of what they already know to have happened…In fact, each emotion is appropriate-logically, and not just morally appropriate-only to certain restricted objects (192).

Kenny (1963) was the first emotion theorist to clearly distinguish between two aspects of the intentionality of the emotions. On the one hand, he argued, emotions have *material objects*, i.e. what Broad had called *epistemic objects*. For example, the material object of Donald Trump's pride may be constituted by the Trump Tower in New York City. But Kenny (1963) also added that "each emotion is appropriate-logically, and not just morally appropriate-only to certain restricted objects" (192). In this case, we are no longer referring to material objects, since the relation between an emotion and its material object holds contingently, rather than with logical necessity. Different people may be proud with respect to different material objects, or the same person may be proud with respect to different material objects at different times.

What Kenny meant is that it "is not possible…to be…proud of…something which one regards as an evil unmixed with good…[or] to envy something which one believes to belong to oneself, or to feel remorse for something in which one believes one had no part" (193). Besides having material objects, Kenny suggested, emotions have *formal objects*, where the "formal object of Φ-ing is the object under that description which *must* apply to it if it is to be possible to Φ it" (189). What Kenny is referring to here is clearly conceptual possibility, not physical possibility. Kenny is not saying that nobody will be physically capable of envying something which one believes belongs to oneself, but rather that the concept of envy will not be instantiated unless what is envied can be described as something which the emoter does not believe belongs to himself.

The same entity, say the Trump Tower, may be the material object of many different emotions for different people. Different people can be angry about it, or envious about it, or afraid about it. Kenny's key point is that, whenever one of these emotions is instantiated with respect to the Trump Tower, it *must* be possible to describe the tower in a way that is logically appropriate to the emotion at hand. For example, if only what has property P is such that it can be envied, then the Trump Toward must be describable as P, which would specify the formal object of envy. The same principle can be applied to every emotion E, in the sense that if only what is P_E can be E-ed, then the description "thing which is P_E" will give us the formal object of E-ing.

By assuming that emotions are essentially feelings, Kenny complained, emotion theorists of the past have failed to account for the logical relation between emotions and their formal objects.

On the one hand, Kenny doubted that feelings can have material objects in the first place. On the other hand, even if they did, "there would be no logical restrictions on the type of [formal] object which each emotion could have", contrary to the assumption that "each emotion is appropriate-logically, and not just morally appropriate-only to certain restricted objects" (192). At best, feeling theorists such as Descartes and Hume understood that an emotion could be about something contingently. What they did not realize, according to Kenny, is that there is something an emotion must be about necessarily. This, in a nutshell, is the *Argument from Intentionality* against the feeling theory.

We can find versions of this argument in basically all early and contemporary cognitivist texts, for example in Broad (1954), Pitcher (1956), Solomon (1976), Nussbaum (2001), Gordon (1987), Greenspan (1988) and many others. The question is: Is it a good argument? I see two main problems with it. The first problem is that sometimes emotions appear not to have epistemological or material objects at all. Often people are not envious or angry about material objects such as a tower or a bear. For example, one may be afraid that the world will come to an end, or ashamed that income is unfairly distributed. What would the material objects be in such cases? We may deal with this sort of difficulty by substituting talk of *material objects* with talk of *particular objects*, where the particular object of E comprises whatever it is that E is contingently about. The particular object could be a physical object, an event, a state of affairs, and the like.

A more resilient difficulty is that even an expanded notion of particular object seems not to accommodate all cases of emotion. On some occasions, it seems that there is really nothing an emotion is contingently about. Sometimes one is anxious, or afraid, or depressed, or angry, but not about anything in particular. These are instances of what we commonly refer to as *objectless emotions*. Their existence creates a problem for Kenny's argument, which relies on the presupposition that "[i]t is possible to be hungry, without being hungry for anything in particular, as it is not possible to be ashamed without being shamed of anything in particular" (Kenny 1963, 60).

The problem is that if the strategy is arguing that emotions cannot be essentially feelings because they are logically appropriate only to certain restricted formal objects, and the formal object of Φ-ing is the material object of Φ-ing "under that description which *must* apply to [them] if it is to be possible to Φ it" (189), the existence of objectless emotions threatens the

Argument from Intentionality at its core. Various strategies have been employed to reconcile objectless emotions with the claim that all emotions have intentionality. One is to say that objectless emotions, contrary to appearances, do have particular objects, just extremely peculiar ones. When we are depressed or angry without an object, what we are depressed or angry about are in effect "things in general", "the blackness of things", "one's present total environment". Another strategy could be to say that objectless emotions are not really emotions, but, say, moods.

A further option, the one favoured by Kenny, is to say that objectless emotions are derivative, and not sufficiently common to worry about them too much. All these strategies seem to me ad hoc, in a way that reveals a philosophically anaemic understanding of the intentionality of the emotions. There are other ways in which we can try to get a grip on intentionality other than in terms of the possession of objects.

The most promising one appears to me that in terms of *norm-answerability*: an X is *intentional* insofar as it is answerable to *norms* establishing how X ought to be. Another way to put the same point is to say that X is intentional insofar as it can be both constitutively *successful* and constitutively *unsuccessful*, namely in agreement or in contrast with its intentionality-constitutive goal. This is the approach to intentionality championed for example by John Searle (1983), who understood intentional states as states having *conditions of satisfaction,* which are the conditions which must obtain in order for the state to be the way it is supposed to be.

This approach to intentionality has originated some of the most influential attempts to naturalize intentionality, such as those proposed by Ruth Millikan (1984, 2004) and Fred Dretske (1986, 1988) in the context of the research program that goes under the banner of teleosemantics. If we think of the intentionality of the emotions in terms of their having particular objects, we run into the obstacle of objectless emotions. Removing that obstacle within the object-based approach to intentionality leads to concocting outlandish objects for objectless emotions, or to dismissing them in some way or other. But nothing is in my view achieved by saying that objectless anxiety is anxiety about the blackness of things, or that anxiety is just a mood, or that we should not worry about it because we are generally anxious about specific things. If we think of the intentionality of the emotions in terms of their conditions of satisfaction, on the other hand, we are led to ask potentially fruitful questions. Is objectless anxiety a case of emotion

which fails to fulfill its intentionality-constitutive purpose? Or is there an intentionality-constitutive purpose which objectless anxiety serves precisely by being objectless?

To answer questions of this sort, we would have to know something about the purposes of emotions, a topic I will address in a systematic form only in chapter 10. For now, let us limit ourselves to transforming the notion of *formal object* from an object-based to a norm-based understanding of intentionality. According to Kenny, the *formal object* of Φ-ing is the particular object of Φ-ing under that description which *must* apply to it if it is to be possible to Φ it.

Under the version I propose, the *formal object* of Φ-ing is a description of its *condition of satisfaction*, namely a description of the condition which must obtain in order for Φ-ing to be the way it is supposed to be. The relation between Φ-ing and its formal object is still logical, but the normative dimension has now been brought center stage. We can now say that, in order for something to be a *correct* Φ-ing, the formal object of Φ-ing *must* be instantiated. We have in effect characterized Φ-ing as the sort of thing which must have a certain constitutive purpose to be what it is. Under this view, whether or not an emotion has a particular object is no longer the key issue. The key issue is whether or not an emotion fulfils its constitutive purpose.

I will assume that the Argument from Intentionality has thereby been defended from the objection that some emotions do not have particular objects. But we now have a new problem in our hands. Kenny and the cognitivists have used the Argument from Intentionality as a weapon against feeling theorists, arguing that "if the emotions were internal impressions…there would be no logical restrictions on the type of object which each emotion could have". When formal objects are understood in terms of conditions of satisfaction, the argument takes the following form: emotions could not be essentially feelings because feelings lack conditions of satisfaction. Kenny's (1963, 14) Argument from Intentionality rested on the assumption that "emotions, unlike pain, have objects: we are afraid of things, angry with people, ashamed that we have done such-and-such".

But once we make room for the possibility that emotions have formal objects despite lacking particular objects, this possibility extends to feelings. Insofar as pain has a formal object in the modified sense I described, namely conditions of satisfaction, then pain can definitely have "logical restrictions" on what the object of its correct instantiation is. Once we think of pain this way, it appears not too hard to imagine what the constitutive goal of pain could be. As argued by Tye (1996), pain may have the constitutive goal of detecting tissue damage. It's not important to

evaluate whether or not this is a good account of pain. The point is that the question of whether or not feelings can have intentionality becomes an open one, once we abandon a restrictive understanding of intentionality in terms of the possession of particular objects.

4.3. THE ARGUMENT FROM DIFFERENTIATION

Another central cognitivist argument against the feeling theory is that differences between feelings, whether we understand them as sensations of pain and pleasure as recommended by Hume or as perceptions of bodily changes as recommended by James, do not map the differences between emotions. For example, two different emotions may have the same feeling attached to them, and the same emotion may have different feelings at different times. This sort of critique, as I anticipated in the section on Skinner (1953), had become very influential after Cannon (1929)'s sweeping attack to the James -Lange theory. Writes Cannon:

> [T]he sympathetic system goes into action as a unit-there may be minor variations as, for example, the presence or absence of sweating, but in the main features integration is characteristic. [The same visceral changes] occur in such readily distinguishable emotional states as fear and rage…[as well as in] such relatively mild affective states as those attending chilliness, hypoglycemia and difficult respirations, and such a markedly different experience as that attending the onset of fever. The responses in the viscera seem too uniform to offer a satisfactory means of distinguishing emotions which are very different in subjective quality (Cannon 1929, 351-352)

The idea here is that even in those cases in which an emotion *does* involve a bodily change, such bodily change will be undifferentiated between different emotion types. In other words, it will not be the case that each emotion has its own specific bodily signature. The Cannonian critique was echoed in practically *all* early cognitive critiques of the feeling theory, and it is to this day one of the main reasons why cognitivists urge the abandonment of the feeling theory as they understand it (e.g. Solomon 2003, Nussbaum 2001). Kenny (1963) stated for example that "[I]t soon became clear that many of the somatic phenomena characteristic of particular emotions occurred also in connection with quite different emotions" (38-39). Bedford (1957) presented the following example: "Indignation and annoyance are two different emotions; but, to

judge from my own case, the feelings that accompany indignation appear to differ little, if at all, from those that accompany annoyance. I certainly find no feeling, or class of feelings, that marks off indignation from annoyance, and enables me to distinguish them from one another. The distinction is of a different *sort* from this" (79). The cognitivist's suggestion was that what differentiates one emotion from the other is the appraisal associated with it. This line of thought was bolstered by two influential experiments published in the 1960s.

The first was performed by Speisman, Lazarus et al. (1964), who demonstrated that subjects exposed to a film depicting what looked like a painful ritual operation on the genitalia of the young members of an African tribe found the movie stressful to different degrees depending on its commentary. A *trauma* commentary, which emphasized the pain of the subjects involved, was associated with more stress than a *denial* commentary, which suggested that no pain was involved, which it turn was associated with more stress than an *intellectualization* commentary, which encouraged subjects to take a detached attitude on the events portrayed. Lazarus took this experiment to show that between a stressor event and the experience of stress there must be an intervening process, which in this case was assumed to be *that which* the commentary allowed to manipulate experimentally.

Lazarus provided a general description of this intervening process in a number of 1950s papers, speaking about stress depending on "differences in the meaning of the situation" (Lazarus, Deese and Osler 1952, 294), or in the "degree of relevance of the situation to the emotive state" (Lazarus and Baker 1956a, 23) or to the "subject's definition of the situation" (Lazarus and Baker 1956b, 267).

In 1960, Magda Arnold (1960) introduced the term "appraisal" to designate the evaluation that brings about the emotion:

> To arouse an emotion, the object must be appraised as affecting me in some way, affecting me personally as an individual with my particular experience and my particular aims (171)

Previous theories of the emotions, Arnold claimed, had mainly focuses on clarifying the causal relation between bodily changes and the experience of emotion, especially in the aftermath of James' controversial claim that the bodily change is, contrary to common sense, what causes the emotional experience. What only a few theories had dealt with, she argued, was "the problem of how cold perception can cause either the felt emotion or the bodily upset" (93).

Further evidence on the crucial role played by appraisal in the elicitation of emotion was offered by Stanley Schachter and Jerome E. Singer (1962). The experimenters' starting point was precisely the Argument from Differentiation against James' feeling theory. They argued that "the variety of emotion, mood, and feeling states are by no means matched by an equal variety of visceral patterns" (379). The absence of physiological differentiation raised the question of what distinguished from one another emotions which were associated with undistinguishable physiological changes. Schachter and Singer reported that dissatisfaction with the Jamesian approach to emotional differentiation led a number of researchers "to suggest that cognitive factors may be major determinants of emotional states" (379). The theory Schachter and Singer wanted to put to the test was what has come to be known as the *two-factor theory of emotions* (a.k.a. cognition-arousal theory of emotions). The theory is summarized in the following passage:

> [A]n emotional state may be considered a function of a state of physiological arousal and of a cognition appropriate to this state of arousal…It is the cognition which determines whether the state of physiological arousal will be labeled as "anger", "joy", "fear", of whatever" (380)

Schachter and Singer set out to demonstrate two main theses. On the one hand, that neither physiological arousal nor cognition (or appraisal) alone can bring about an emotion. On the other hand, that when cognition and physiological arousal occur jointly, cognition is what determines the identity of the emotion, whereas physiological arousal is what gives emotionality to the experience.

The basic experimental setting consisted of injecting epinephrine in groups of emotional subjects, so as to artificially produce autonomic arousal, and then leave them in a room for 20 minutes together with a stooge playing the role of another experimental subject participating to the test, officially described as a vision test. During these 20 minutes, subjects were submitted to either of two sets of experimental conditions, meant to elicit respectively euphoria and anger. Shachter and Singer claimed to have observed that subjects indeed got euphoric in the euphoria condition and angry in the anger condition, more so if they did not know that they had been injected epinephrine, indicating that the presence of arousal generates a need to cognitively label it. They also claimed that the subjects who had not received an injection did not get angry and euphoric as much as the subjects injected, concluding both that physiological arousal is

necessary for the instantiation of emotion and that the way it is cognized determines which emotion is instantiated.

The experiment, however, is fraught by methodological flaws, concerning for example the unreliability of the measures used to detect the presence and intensity of anger and euphoria and the fact that subjects were asked to self-rate their emotions after the effects of epinephrine had dissipated. Moreover, the data fail to fully support the experiments' conclusion, in the sense that the alleged differences between emotional indexes in different conditions are often statistically insignificant, and subjects who are not administered epinephrine do experience anger and euphoria in the two experimental settings. Moreover, the experiment has turned out to be difficult to replicate. Despite these shortcomings, the main message of Schachter and Singer's experiment, namely that cognitions or appraisals are necessary to establish the identity of the emotions, has been taken to heart by psychologists ever since. It is no exaggeration to say that the experiment represented the main propellant of cognitivism in psychology in the 1960s and 1970s.

The main problem with the Argument from Differentiation is that, as I see it, it is perfectly compatible with the feeling theory as understood by its proponents. Descartes and Hume would have readily agreed with the point that the quality of emotional experiences is not sufficient to differentiate them. I reported Descartes' (1650) remark that to simply assume that a passion is an agitation whose proximal cause is the movement of animal spirits "does not enable us to distinguish between the various passions: for that, we must investigate their origins and examine their first causes" (art. 51). I also mentioned that Hume described the indirect passions as deriving from a *double relation* of *impressions* of pain and pleasure, which will be common between different emotions, and *ideas*, which will serve to distinguish between emotions characterized by the same impressions. James (1884, 1990), the main target of the Argument from Differentiation, did suggest that emotions may differ in terms of their physiological underpinnings. What is rarely noticed is that he did not commit to this hypothesis, only presenting it as "abstractly possible". James (1884) wrote:

> The various permutations and combinations of which … organic activities are susceptible make it *abstractly possible* that no shade of emotion, however slight, should be without a bodily reverberation as unique, when taken in its totality, as is the mental mood itself (192)

James did not develop this point, and did not rely on it in his Argument from Conceivability (see subsection 2.3.1). It is compatible with the Jamesian proposal to say that emotions do not differ from one another physiologically, yet amount to perceptions of bodily changes. This is because it is open to a Jamesian to say that emotions are identified not only in terms of their bodily changes but also of the appraisals associated with them. This turns out to be the strategy adopted by Prinz (2004a) in his recent defense of a James-inspired theory of emotions, which I discuss in chapter 7. It must also be said that since Cannon wrote on the topic, it has become less clear whether or not different emotions differ in terms of bodily changes, if and when they involve them. I will further discuss this topic in chapter 10.

4.4. CONCLUSION

In this chapter, I described the main arguments which led to the demise of the feeling theory in the early 1960s and 1970s. What came after was the cognitivist theory of emotions.

The Argument from Absent Consciousness made a valuable point, namely that we should not think of the emotions exclusively in terms of the special states of consciousness associated to them. Some emotions lack A-consciousness, in the sense that they occur but emoters cannot access the thought that they do. Some other emotions lack the specific type of bodily P-consciousness commonly associated to them.

The Argument from Intentionality under the interpretation I have offered made an important point, namely that the emotions have constitutive conditions of appropriateness. However, I have argued that this argument alone would not prevent them from being essentially feelings.

The Argument from Differentiation, finally, held that since feelings do not differ from one another enough, something else is needed to differentiate emotions from one another. The cognitivist proposal is that appraisal is what does the differentiating.

5. EMOTIONS AS ADAPTATIONS

The approaches I have explored so far tried to explain what the emotions are without focusing primarily on their origin and function. The traditions I explore in this chapter and the next started instead from the presupposition that understanding emotions is understanding what sorts of problems they are meant to solve.

According to the evolutionary tradition, an emotion is an adaptive solution to a fundamental life task. The basic tenet of the evolutionary tradition can also be understood as a generalization assumed to be true of all emotions, namely that they were selected for in the ancestral past by virtue of their beneficial impact on reproductive fitness. One of the notable features of the evolutionary approach is that it calls our attention on the communicative function of the emotions, all but neglected by the traditions I have surveyed so far.

Charles Darwin (1872) was the first to study emotional expressions within an evolutionary framework. Notably, Darwin did not argue that the emotions evolved because of their communicative function. However, he laid the groundwork for a defense of this claim, and more importantly, for an understanding of the emotions as a whole as adaptive traits. The evolutionary approach was revived by Silvan Tomkins (1962, 1995) and Robert Plutchick (1962, 1970, 1980), and it came to maturity with Paul Ekman (1969, 1971, 1987, 1992, 1999a, 1999b) and Carrol Izard's (1969, 1971, 1977, 1980, 1992, 1993).

5.1. EMOTIONS AS SOLUTIONS TO FUNDAMENTAL LIFE TASKS

5.1.1. Darwin

In *The Expression of the Emotions in Man and Animals* (1872), Darwin did not offer a theory of the emotions, but rather a theory of emotional expressions, aiming to shed light on their origin and nature. By *emotional expression*, Darwin meant "movements or changes in any part of the body", for example "the wagging of a dog's tail, the drawing back of a horse's ears, the shrugging of a man's shoulders, or the dilatation of the capillary vessels of the skin" (28).[6] The main Darwinian novelty was to argue that the understanding of human expressions demanded an evolutionary framework and an appreciation of their continuity with animal expressions. As he put it, "[w]ith mankind some expressions, such as the bristling of the hair under the influence of extreme terror, or the uncovering of the teeth under that of furious rage, can hardly be understood, except on the belief that man once existed in a much lower and animal-like condition" (12).

The Darwinian account, however, did not fully exploit the potential of the evolutionary approach with respect to the emotions. On the one hand, it only focused on emotional expressions, rather than on emotions as a whole. On the other hand, and more surprisingly, Darwin did not argue for the evolutionary function played by expressions, explicitly denying that the communication of information contributed to explaining their origin. Despite these limitations, Darwin was the first to offer an evolutionary framework for the understanding of the emotions, a framework whose potential was explored by others in the second half of the 20th century. To understand the Darwinian account of emotional expressions, it is important to understand how it fit into Darwin's general research agenda. The book, initially intended as a section of *The Descent of Man* (1871), had the objective of showing that the comparison between human and animal expressions offers further support to the hypothesis of evolution by natural selection, the revolutionary idea Darwin had proposed in his masterwork, *On the Origin of Species* (1859). This is the conclusion Darwin explicitly drew at the end of his book on expressions, when he wrote that "the study of the theory of expression confirms to a certain limited extent the conclusion that man is derived from some lower animal form".

[6] All page numbers in this section are referred to an electronic copy of Darwin's book edited by van Wyhe and accessible at http://pages.britishlibrary.net/charles.darwin/.

To achieve the objective of using emotional expressions to support evolution, Darwin had to deal with a tradition of research which understood them under the assumption that “species, man of course included, came into existence in their present condition” (10). Authors working on expression in this tradition included most prominently Bell (1844), the researcher who according to Darwin “laid the foundations of the subject [of expression] as a branch of science” (2), as well as Duchenne, Gratiolet, and Piderit. The only researcher on expressions to apply the evolutionary approach was Herbert Spencer, whose account Darwin developed. Bell, Darwin’s chief target, had stated that “many of our facial muscles are "purely instrumental in expression;" or are "a special provision" for this sole object.””. (Bell, as quoted in Darwin (1872, 10)). Bell’s (1844) view was that God had given facial muscles to human beings so as to allow them to communicate their inner feelings.

Ekman (1997, xxxiv), in his preface to Darwin (1872/1997), suggested that Bell’s emphasis on communication may explain why Darwin did not put the evolutionary value of communication center stage in his account of emotional expressions. This strikes me as a sensible suggestion, even though Ekman does not explain what he took Darwin’s implicit reasoning to be. A reasonable interpretation seems to me the following. Darwin was primarily interested in arguing that expressions had not been given to men by God, but had rather resulted from evolution. To make this point, Darwin denied that expressions have primarily purposes of communication, as this was Bell’s rationale for assuming that God had given them to men. Supporting the evolutionary alternative on the basis of this denial demanded explaining how evolution had given expressions to men without invoking their communicative function. Darwin could have of course chosen another path, namely conceding to Bell that human emotions were there because of their communicative function, but denying the other half of his thesis, and show that evolution, rather than God, had been the granting authority.

In the following passage, we can appreciate the Darwinian strategy at work. Commenting on Bell’s claim that God had given facial muscles to men in order to allow them to communicate their emotions, Darwin commented that “the simple fact that the anthropoid apes possess the same facial muscles as we do, renders it very improbable that these muscles in our case serve exclusively for expression” (10). Darwin’s strategy was to capitalize on the fact that “no one, I presume, would be inclined to admit that monkeys have been endowed with special muscles solely for exhibiting their hideous grimaces” (10) and argue that if our expressions derive from

those of monkeys, then we should conclude that we have not been endowed with them *by God* solely for exhibiting our own grimaces. Darwin's central explanatory principle was a development of a view first expounded by an author whose identity, Darwin says, "I have not been able to ascertain" and opposed by Bell, namely that "what are called the external signs of passion, are only the concomitants of those voluntary movements which the structure renders necessary" (Bell, as quoted by Darwin (1872, 9)).

This view, further articulated by Spencer against Bell, became Darwin's trademark principle of "serviceable associated habits". According to such principle, some emotional expressions are involuntary vestiges of voluntary actions that used to be serviceable in the ancestral past and kept being associated by force of habit or by reflex to the states of mind that brought them about. For example, the snarling expression exhibited by a human being in wrath, which consists of "[u]ncovering the canine tooth on one side", "is the same as that of a snarling dog" and "[i]t reveals his animal descent; for no one, even if rolling on the ground in a deadly grapple with an enemy, and attempting to bite him, would try to use his canine teeth more than his other teeth" (253). The snarling expression is for Darwin inherited from "our semi-human progenitors", who "uncovered their canine teeth when prepared for battle, as we still do when feeling ferocious…without any intention of making a real attack with our teeth" (253). Once such expressive habit has been established, other expressions are generated on its basis through the subsidiary principle of "antithesis".

According to such principle, states of mind opposed to those eliciting expressions according to the principle of serviceable associated habits will recruit expressions directly in antithesis to them. As Darwin remarks, if a dog in a ferocious state of mind displays a fixed stare, walks tall and holds his tail stiff and upright, a dog in a placid state of mind will not look intently, will almost crouch, and will lower and wag the tail. These two principles, however, cannot explain all expressions, in the sense that some of them do not appear to be associated with, or opposed to, any action that used to be serviceable. Darwin's principle of the "direct action of the nervous system" was introduced to take care of such cases.

According to such principle, some emotional expressions are the direct result of the excitation of the nervous system. "When the sensorium is strongly excited", Darwin wrote, "nerve-force is generated in excess" (29), and expressions result directly. Darwin cited the examples of "trembling of the muscles, the sweating of the skin, the modified secretions of the

alimentary canal and glands" (68). The three principles operated in concert, and Darwin acknowledged that it is often hard to apportion their individual influences. A question we must get clear about is: Did Darwin deny that emotional expressions communicate and play a useful role for human beings?

The answer is no, since Darwin explicitly acknowledged the communicative usefulness of expressions. For example, he stated that "[t]he movements of expression in the face and body, whatever their origin may have been, are in themselves of much importance for our welfare" (365). A mother and an infant, Darwin remarked, usefully communicate through facial expressions. We perceive other people's sympathy towards us "by their expression; our sufferings are thus mitigated and our pleasures increased" (365). Generally speaking, expressions "reveal the thoughts and intentions of others more truly than do words, which may be falsified" (365). Darwin's point is rather that the origin of expressions is not due to their usefulness for us, in the sense that "every true or inherited movement of expression seems to have had some natural and independent origin" (356).

To put is in modern terms, Darwin's view was that emotional expressions are at best *exaptations*, namely traits which currently play a role other than the one they were selected for. Darwin's point was that although expressions "often reveal the state of the mind, this result was not at first either intended or expected", as they have "been at first either of some direct use [as a serviceable action], or the indirect effect of the excited state of the sensorium" (357). Darwin offered further evidence to the effect that expressions are inherited in humans. "That these and some other gestures are inherited", he stated, "we may infer from their being performed by very young children, by those born blind, and by the most widely distinct races of man" (353). If emotional expressions are the same prior to learning (in children), in the absence of learning (in the blind) and despite differences in learning (in different races), then Darwin thought the case for their inherited status would be strengthened.

The study of emotional expressions, however, did not have for Darwin exclusively the purpose of bolstering the hypothesis of evolution. It also had the purpose of clarifying "how far particular movements of the features and gestures are really expressive of certain states of the mind" (13), an issue of independent interest to him. To clarify what expressions stand for what states of mind, Darwin combined the study of animals with that of infants and the insane, both assumed to be very expressive, and that of works of art, which disappointed him because he

believed the pursuit of beauty had prevented artists from portraying intense emotional expressions.

His most interesting contributions on the study of the "state of mind-expression" relation were the picture and questionnaire techniques he pioneered. The former consisted of showing people pictures of emotional expressions, and ask them what emotions they expressed. Darwin only applied it to English people, but his technique has become the most used one for the study of the universality of emotional expressions in different cultures. Darwin's own cross cultural technique consisted of sending questionnaires to English observers familiar with people from Africa, America, Australia, Borneo, China, India, Malaysia and New Zealand, and ask them about the expressions of the natives. Darwin's questions were poorly formulated, because they conveyed Darwin's expected answer (e.g. "Is astonishment expressed by the eyes and mouth being opened wide, and by the eyebrows being raised?"). Moreover, Darwin did not collect a sufficiently large sample of answers (only 36 questionnaires were returned to him). The data he collected, however, convinced him that "the same state of mind is expressed throughout the world with remarkable uniformity".

I take this to be a formulation of what I will call the *universality of emotional expression thesis*. According to such thesis, the emotions are expressed in the same way in all cultures in which they exist, and at least some emotions exist in all cultures. If the thesis is true, then we should be able to find *some* emotions expressed in the same way in all cultures. The thesis can be strengthened by adding to it that all emotions exist in all cultures (*universality of emotion thesis*) and that all emotions have their own *distinctive* expression in all cultures in which they exist (*distinctiveness of emotional expression thesis*). If the three theses are fulfilled, then all emotions are present in all cultures and they are expressed in the same, distinctive way. The evolutionary lesson Darwin drew from the cross-cultural evidence he collected was that all races of men must have come from a common ancestry, because they share some emotions and express them in the same way.

Among the emotions present and expressed in the same way everywhere, Darwin listed anger, indignation, contempt, disgust, scorn, disdain, shame, and "good spirit". Darwin was doubtful instead that emotions such as "Jealousy, Envy, Avarice, Revenge, Suspicion, Deceit, Slyness, Guilt, Vanity, Conceit, Ambition, Pride, Humility" (262) have distinctive expressions reliably associated to them, and did not offer data in support of the view that they existed in all

cultures. As argued by Ekman in his introduction to Darwin (1872/1997), after an initially successful run, Darwin's book on expression "became virtually forgotten for 90 years". According to Ekman, one of the chief reasons for this neglect was that Darwin proposed an understanding of the emotions as *unlearned* and *biologically determined*, a view which contrasted with the research programs that dominated the first half of the 20th century.

The behaviorists (see chapter 3) resented the unlearned nature of Darwinian emotions, as they took learning through conditioning to be the tool by means of which every behavior, including emotional ones, could be shaped. We can add to this the fact that Darwin endorsed, following Spencer, a theory of emotions as feelings, which the behaviorists strongly opposed. The cultural relativists resented Darwin's emphasis on biological nature, arguing that the emotions were more a matter of nurture than a matter of nature. In particular, as we shall see in subsection 5.2.4, starting in the 1920s Darwin's evidence on the universality of emotional expressions was questioned by anthropologists and social scientists.

By the early 1960s, the cultural relativists seemed to have won the day. The renewal of the evolutionary tradition occurred mainly through Silvan Tomkins' *Affect, Imagery and Consciousness* (1962) and Robert Plutchick's *The Emotions: Facts, Theories and a New Model* (1962). The insights of these two books were developed in Ekman and Izard's theory of basic emotions. It should also be pointed out that in the 1960s Konrad Lorenz singled out Darwin's work on expressions as sharing the founding insight of ethology, which he ascribed to Charles Otis Whitman and Oskar Heinroth. As Lorenz (1965) wrote in his preface to Darwin's *The Expression of Emotion in Animals and Humans*, "reading between the lines" it becomes clear that Darwin was aware that "behavior patterns are just as conservatively and reliably characters of species as are the forms of bones, teeth, or any other bodily structure" (xii).

Once the motor pattern of biting has been selected for, Lorenz (1965) remarked, it remains on board even when biting is no longer of use, just as bodily structures do when they have been selected for. And as bodily structures such as gill slits can take on a new function for which they were not selected – e.g. the function of ears -, so the motor behavior of biting, abridged into a "snarling" movement, can take on the new function of communicating.

5.1.2. Tomkins

Silvan Tomkins' (1962, 1995) theory of emotions, which he relabeled *affects*, must be understood as resulting from two main theoretical objectives. The first objective was to argue that affects, rather than drives, constituted the "primary innate biological motivating mechanism". This objective demanded describing how affects differ from drives, and why they are primary. The second objective was to argue that affects are discrete entities, rather than points along a continuum of variation as recommended by dimensional theories of emotion. Tomkin's solution was to describe affects as differing from drives because of their being abstract, general and urgent, and as differing from one another because there are perceptions of distinct *facial changes* governed by an evolved affect program. Let us briefly characterize the *drive theory* of motivation and the *dimensional theory* of emotions from the perceived shortcomings of which Tomkins developed his own theory of affects.

The concept of *drive* was coined in 1918 by Robert S. Woodworth, as a substitute for the older *instinct*. Although the notion of drive comes in different flavors, the basic idea is that drives motivate behaviors by virtue of internal self-regulation of physiological imbalances which generate a *need* to be satisfied. Cannon (1929) described as *homeostasis* the self-regulating biological mechanism allowing the detection and stabilization of physiological imbalances. As Tomkins (1995) put it, "for some few thousand years, up to and including Hull and Freud", the answer to the question of what motivates organisms was that "the human animal, is driven the breath, to eat, to drink, and to engage in sex" (101). In Freud's case, drives reside in the Id and are unconscious and irrational, a feature absent from the behaviorist theory of drives developed by Clark Hull in the 1940s.

The *dimensional theory* of emotion stems from an old idea, namely that feelings differ from one another in terms of their pleasure and pain and in terms of their intensity. In 1896, Wundt distinguished three dimensions of variations for feelings - pleasure-displeasure, excitement-calm and strain-relaxation – and argued that every distinction between feelings could be captured along these three dimensions. Dimensional theories of the emotions attempt to shed light on emotion categories by associating them to points in multidimensional spaces individuated by a few dimensions of quantifiable variation. Several lists of such dimensions have been offered since Wundt (1896), but the most popular three have been the dimensions of *pleasure* (or evaluation, valence, positivity), *activation* (or arousal or activity) and *potency* (or

power, control, dominance). For example, anger could be represented on a continuum of variation by high displeasure, high arousal, and high dominance.

The combined effect of the drive theory of motivation and of the dimensional theory of emotion was that specific affects - fear, anger, sadness, happiness, disgust - were not the primary object of study, whereas drives, pleasure, activation, and potency received all the theoretical attention. Tomkin's aimed to change what he judged to be a centuries-long neglect of affects. Tomkins' (1995) first move was to argue that drives are powerless as motivators in the absence of affects, and differ from them in a number of ways:

> [T]the drive must be assisted by affect as an *amplifier* if it is to work at all …. The affect is, therefore, the primary motivational system because without its amplification, nothing else matters, and with its amplification, anything else *can* matter. It thus combines urgency and generality. It lends power to memory, to perception, to thought, and to action no less than to the drives (355–356)

Consider gasping for breath in a case of anoxia, or getting excited at the sight of an attractive woman or man. Traditionally, these would have been considered paradigmatic examples of how the drive to breath or the drive to engage in sex are powerful motivators. Tomkins' point is that the drive signals – lack of oxygen, presence of sexual object – would not motivate unless they recruited the affects of, respectively, fear and excitement. Conversely, once such affects are recruited by things other than drives, motivation ensues. In other words, drives are, differently from affects, neither sufficient nor necessary for motivation, and consequently not primary. For example, we can be afraid or excited because of an act of cognition, a notion Tomkins associated to propositional thought. One can get excited at the thought that a certain mathematical solution is very elegant, or afraid at the thought that nuclear war may one day wipe out civilization. As Tomkins (1995) puts it with respect to excitement, "[a]lthough mathematics and sexuality are different, the excitement that amplifies either cognitive activity or drive is identical" (53).

They way in which affects amplify, Tomkins remarks in the passage above, is by combining *urgency* and *generality*. Affects make one care about the drive signal or the perception or the thought *amplified* by it in a very powerful and insistent way, in the sense that they make coping with the arousing source a priority. The idea that emotions can be characterized by their urgency is be the central idea around which my own theory of emotions is constructed. I will come back to what emotional urgency is, and why it is important in chapter 10.

Besides being *urgent*, Tomkins (1995) takes affects to be *general*. Most importantly, differently from drives affects have *generality of time* and *generality of object*. They have *generality of time* in the sense that they neither occur in cycles, nor at a pre-determined time nor for a fixed duration. One is not afraid in rhythmic cycles, at particular times and for a pre-established amount of time. The hunger or thirst drives, on the other hand, occur in cycles, at times in which resources are depleted and for the time needed to achieve satiation. Affects have *generality of object* in the sense that what one can love or be afraid of anything, and consequently be motivated to all kinds of responses. On the other hand, the hunger or pain drives motivate a more narrow range of responses, namely eating responses or responses addressed to taking care of the wound. After having established that affects are the primary motivators, that they are urgent and that they are general, Tomkins' proceeded to explain how many affects there are.

Tomkins' (1995) view was that "affects are primarily facial behaviors and secondarily outer skeletal and inner visceral behavior" (217), in the sense that visceral changes slowly follow facial ones and that "[w]hen we become aware of these facial responses (with or without concurrent visceral responses), we are aware of our affects" (217). Under this view, the reason why an affect motivates is that there is motivating feedback from facial and visceral responses. Such responses are organized by a subcortical program – an affect program - which Tomkins assumed to have evolved. Since affect programs control facial and visceral changes, affects are type-identified by unique sets of such changes. Writes Tomkins (1995):

> If each innate affect is controlled by inherited programs that in turn control facial muscle responses, autonomic blood flow, respiratory, and vocal responses, then these correlated sets of responses will define the number and specific types of primary affects (58).

Tomkins (1995) argued that there are nine "discriminable distinct sets of facial, vocal, respiratory, skin and muscle responses", namely "interest, enjoyment, surprise, fear, anger, distress, shame, contempt and disgust" (58). For example, Tomkins described "fear-terror" as "eyes frozen open, pale, cold, sweaty, facial trembling, with hair erect", "shame-humiliation" as "eyes down, head down", "anger-rage" as "frown, clenched jaw, eyes narrowed, red face". What is absent from these definitions of emotions are all marks of emotionality other than expressive and physiological ones. In effect, Tomkins understood the feeling of affect roughly as James did, with the substitution of visceral changes with facial changes as what type-identifies affects.

I argue that the reason why Tomkins argued that cognitions and behaviors are not characteristics of affects as such is that he wanted to characterize affects as an independent motivational mechanism. He was struck by the *generality* of affect, namely by the variety of cognitions (or drives) that can cause them and the variety of behaviors (or mental acts) they can cause. What Tomkins did not consider is that there is a level of description that may allow us to characterize affects in terms of their appraisals and behaviors despite their generality. For example, one can be afraid of many things, but, at first blush at least, they all appear to qualify as appraisals of danger. Similarly, one can do many things when in fear, but, at first blush at least, they all qualify as avoidance behaviors. The differentiation of the emotions from one another, as well as the account of their intentionality, demands that we go beyond the feelings of emotions, whether they are understood as feeling of visceral changes or feelings of facial changes.

Tomkins' (1962) central assumption was that each of the nine primary affects had distinct facial expressions associated to them. However, he did not have experimental data to back this up. The question has since become very controversial.

What appears to be clear is that at least some of the things we ordinarily call emotions do not have a distinctive facial signature. This possibility was known to Tomkins, who did not include for example guilt and shyness in his list of primary affects. He argued that "[o]ne should not distinguish shame from guilt and shyness as affects, but rather as affect complexes of shame plus varying perceived and conceived causes and consequences" (1995, 61). *Affect complexes* are understood by Tomkins as "complex assemblies of affects and perceived causes and consequences" (59). The problem is that neither guilt nor shyness are kinds of shame with particular causes, differently from shame caused by being naked in public and shame caused by being caught in a strip club by one's girlfriend.

The category of affect complexes is not clearly defined by Tomkins, and no explicit rationale is given for drawing the boundary between *affects* and *affect complexes* in terms of possession of distinctive facial expressions. Despite these limitations, I want to argue that Tomkins' affect theory was a significant step forward in the history of emotion theory, because it offered a framework for the evolutionary understanding of emotions. Tomkins chose the face as the primary site of affect in part because he believed the communicative function of emotions contributed to their evolution. He wrote that the human face "seems to have evolved in part as an organ for the maximal transmission of information, to the self and to others, and the information

it transmits is largely concerned with affects" (1995, 218). In this respect, his theory differs from Darwin's (1872) theory, who assumed that human facial expressions are mostly vestiges of serviceable behaviors, currently useful but not evolved for their communicative function.

At the same time, Tomkins believed that the primary biological function of affects was not that they sent signals, but that they were "sources of motivating feed-back" (90). But how to affects motivate? How do they communicate? What is the evidence that they evolved? Answering these questions is a task Tomkins (1962) left to a number of young psychologists who became interested in facial expressions in the early 1960s, most prominently Paul Ekman and Carroll Izard.

5.1.3. Ekman

The inspiring thought at the foundation of Ekman's (1984, 15) version of *affect theory* is that "emotions evolved for their adaptive value in dealing with fundamental life tasks", a feature which allegedly distinguishes them from all other affective phenomena. Generally speaking, a trait is an adaptation in case its existence results from a process of natural selection, and it is adaptive at a particular time t just in case it generates differential fitness (see below). The two properties are orthogonal to one another at any time t, in the sense that a trait that is adaptive at time t may or may not be an adaptation, and a trait that is an adaptation may or may not be adaptive at time t.

Ekman's claim is that emotions are adaptations, without necessarily being currently adaptive. The first emotion theorist to explicitly articulate the idea that emotions evolved not merely because they motivate with urgency (Tomkins' view), but because they motivate to deal with fundamental biological challenges was Plutchick (1962, 1970). Plutchick characterized the primary emotions as those fulfilling "the basic adaptive or prototype functions". Under this view, he claimed that there are 8 primary emotions:

Primary Emotion	***Biological Function***
Fear	Protection
Anger	Destruction
Joy	Reproduction

Sadness	Deprivation
Acceptance	Incorporation
Disgust	Rejection
Anticipation	Exploration
Surprise	Orientation

Figure 3: Plutchick's list of primary emotions

"In order to provide a general definition of emotion" Plutchick (1970) argued, "we need to use the functional or adaptational language". The main virtue of such language is that, differently from the language of, say, feelings, it "is the most general and applies to humans as well as other animals" (12). Under this view, "[a]n emotion is a patterned bodily reaction of either protection, destruction, reproduction, deprivation, incorporation, rejection, exploration or orientation, or some combination of these, which is brought about by a stimulus" (1970, 12). According to Plutchik, emotions are primary or derived from a blend of primary emotions, in the same way in which "all colors can be considered to result from a mixture of just a few primary colors".

Ekman (1999b) developed the idea that emotions must be defined in adaptational language, and referred to the emotions as *basic emotions*, to emphasize that they are understood from an evolutionary viewpoint. But he also added "I do not allow for "non-basic" emotions" (57), a position which implies that nothing can be an emotion unless it emerged from the ancestral past as a solution to a fundamental life task. To illustrate his understanding of life tasks, Ekman cited Tooby and Cosmides's (1990) description, which includes "event times that recurred innumerable times in hominid evolutionary history" such as "[f]ighting, falling in love, escaping predators, confronting sexual infidelity, experiencing a failure-driven loss in status, responding to the death of a family member" (1990, 92). Ekman has changed his views on the characteristics and number of basic emotions through time, and I will only refer to his latest position on the matter.

In Ekman (1999b), he provides a list of eleven characteristics of basic emotions:

1. Distinctive universal signals
2. Distinctive physiology
3. Automatic appraisal, tuned to:

4. Distinctive universals in antecedent events
5. Distinctive appearance developmentally
6. Presence in other primates
7. Quick onset
8. Brief duration
9. Unbidden occurrence
10. Distinctive thoughts, memories images
11. Distinctive subjective experience

He does not "think any of the characteristics should be regarded as the *sine qua non* for emotions, the hallmark which distinguishes emotions from other affective phenomena" (1999b, 47). Although Ekman does not fully clarify the issue, my impression is that he is convinced that, at least at this stage of research, something should count as a basic emotion when "enough" of the characteristics are fulfilled. This will create some differences between basic emotions, but maintain the unifying property that they all evolved. Ekman's choice of characteristics is motivated by central assumptions about the nature of the adaptive value of the emotions. Writes Ekman (1999b):

> Quick onset is central to the adaptive value of emotions, mobilizing us quickly to respond to important events. It is also adaptive for the response changes which can occur so quickly not to last very long unless the emotion is evoked again (54).
>
> I believe it was central to the evolution of emotions that they inform conspecifics, without choice or consideration, about what is occurring (47).

Affective phenomena which are emotions, therefore, are understood by Ekman as having been selected for because of their quick and short-term resource mobilization and communicative effect. This is the main intuition that guides Ekman's research program. If the emotions have been selected for, then they should be present in other primates and have a distinctive appearance developmentally. If the emotions have been selected for in part because of their speed and efficiency, then (a) their appraisal mechanism should be automatic, namely "capable of operating with great speed [and] without awareness" (Ekman 1984, 15) (quick onset), (b) "there will be

some common elements in the contexts in which emotions are found to occur" (distinctive universals in antecedent events), (c) their duration is "certainly…not hours or days, but more in the realm of minutes or seconds" (brief duration) (16) and (d) "there should also be physiological changes preparing the organism to respond differently in different emotional states" (distinctive physiology). If emotions have been selected for in part because of their communicative function, then we should expect them to be endowed with distinctive facial expressions invariant in all cultures (distinctive universal signals).

The unbidden occurrence of basic emotions is a consequence of the automaticity of appraisals, changes in expression and changes in physiology, where the automaticity explains why "we often experience emotions as happening to us". Ekman also mentions that he expects the emotions to "regulate the way in which we think" (distinctive thoughts, memories, images), and that they comprise a distinctive subjective experience, although he distrusts the evidence about such experience because "most of what we know about subjective experience comes from questionnaires, filled out by people who are not having an emotion, trying to remember what it feels like" (Ekman 1999b, 55).

What is notably absent from the list of characteristics is the presence of distinctive behavioral responses, which are obviously characteristic of emotions if they have evolved to deal with specific life tasks (e.g. attack behaviors are characteristic of anger). I suspect the reason why Ekman did not include them in the list is that they appeared to him to be too open-ended for being recruited by an *affect program*, the "mechanism that stores the patterns for [the] complex organized responses" (1980, 82) characteristic of the basic emotions. I already discussed this point with respect to Tomkins, so I won't speculate on this matter further. I will take affect programs to govern behavioral responses as well, a view that is broadly compatible with Ekman's statements on the matter.

How many basic emotions are there? Ekman's (1999b, 55) answer is that "[a]lthough the evidence is certainly not available now", there are fifteen emotions which "will be found to share the characteristics listed" above: amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, pride in achievement, relief, sadness/distress, satisfaction, sensory pleasure, and shame . Each of these emotions represents a theme or a family, which admits of variations. For example, the anger family will include rage, irritation, frustration, fury, the sadness family will include sorrow, melancholy, disappointment, and so on. Ekman (1999b,

55) points out that "[e]ach member of an emotion family shares the characteristics" that distinguish the family from all other families. This implies that however rage and frustration may differ, Ekman assumes them to have the distinctive universal signals of anger, the distinctive physiology of anger, the distinctive quick onset of anger, and so on.

Ekman is aware that there are things we ordinarily call emotions which are not likely to fulfill enough or even any of the characteristics of basic emotions. He considers the examples of interest, romantic love, parental love, hatred, grief, and jealousy, and remarks that he does not expect we will find evidence to the effect that they are basic emotions. The examples could be multiplied, in the sense that many of what James called the "subtler emotions" (e.g. intellectual, aesthetic and moral emotions) do not appear likely to have enough of the characteristics of basic emotions. A further difficulty is that there are tokens of the basic emotion types which appear to lack the characteristics of basic emotion types. For example, there are tokens of anger, fear, disgust, sadness, surprise and joy - the six basic emotions for which Ekman thinks we have the strongest evidence - which lack automatic appraisal, last for a long time, have no distinctive facial signals, have no distinctive physiology, and have no distinctive subjective experience. An unconscious token of fear of failing, for example, lacks all such characteristics.

Ekman's strategy is simply to insist that there are no emotions other than the basic ones. But this clearly calls for a rationale, in the sense that it is not enough to say that jealousy is not an emotion because it lacks the characteristics of basic emotions. The question is: Why should the eleven characteristics of basic emotions *matter*?

To answer this question demands getting clear on the appropriate methodology for the study of emotions, a task I tackle in chapter 9.

5.2. THE ARGUMENT FROM EVOLUTION

5.2.1. The pitfalls of adaptationist thinking

Generally speaking, what counts as evidence that a trait T was selected for is that several empirical facts about it are best explained by the hypothesis that there has been an "adaptive environment" E and an "adaptive time" t in which:

(1) Organisms Os varied as to whether they had trait T (principle of variation)

(2) The offspring of organisms with T had T (principle of heredity)

(3) Having T was beneficial in E in comparison with not having T, in the sense that it gave individuals with T "differential fitness" (ceteris paribus, organisms with T left more offspring than organisms without T) (principle of differential fitness)

Common examples of empirical facts that offer support for an evolutionary explanation of T are the presence of T in similar form in related species, the presence of T in newborn organisms and, in the case of human traits, the presence of T in similar form in different cultures. As Griffiths (1997) has argued, we must be very careful about the way in which we interpret these facts, which are neither necessary nor sufficient for T to be an adaptation. For example, T may be an adaptation but appear in very different form in different species, since "homologies", which are "traits possessed by all and only the descendants of the ancestral species in which these traits originate" comprise "examples which have been radically transformed in form and function", such as "human arms", the "wings of birds" and the "flippers of dolphins".

On the other hand, T may not be an adaptation, but appear in similar form in different cultures, because of species-constant learning. Before describing the evidence for the evolution of the emotions, it is important to point out that evolutionary explanations are often offered in emotion theory without much supporting evidence. A case in point is represented by the account of the emotions proposed by a number of evolutionary psychologists. Evolutionary psychologists view the mind as "a crowded zoo of evolved, domain-specific programs" (Tooby and Cosmides 1990, 91), and aim to apply evolutionary explanations to a much larger set of emotions that those admitted as basic by affect program theorists such as Ekman and Izard. As argued by Griffiths (1997), this is a result of "adaptive thinking", which he takes to characterize the evolutionary psychology program as a whole. The issues surrounding the notion of adaptive thinking are complex, and they cannot be discussed in any detail here.

For the purposes of this dissertation, adaptive thinking is to be understood as the arbitrary formulation of adaptationist hypotheses and their acceptance on the basis of scant evidence, roughly along the lines first denounced by Gould and Lewontin (1978). Let us briefly consider Tooby and Cosmides (1990) account of the emotions. Their view is that "[t]o the extent that

situations are structured and recurred through evolutionary time, their statistical properties can be used as the basis for a special kind of psychological adaptation: an emotion" (410). More precisely, Tooby and Cosmides argue that "conditions or situations relevant to emotions", namely situations that call for emotion-programs as adaptive solutions, are distinguished by five characteristics: they recurred ancestrally, they could not be successfully negotiated in the absence of a superordinate level of program coordination, they had a rich and reliably repeated structure, they had reliable cues signaling their presence, and they were on a type in which an error would have resulted in large fitness costs.

The problem is that it is easy to come up with plausible accounts of recurrent situations of this kind in correspondence with most if not all emotions, and describe these "situations" at an arbitrary level of specificity. Tooby and Cosmides argue for example that they consider "fear of predators", "guilt", and "sexual jealousy" to be adaptations. According to their account, however, we may just as well consider *fear of lions*, or *guilt about incest*, or *sexual jealousy towards a daughter* to be distinct emotion-programs. For example, we may say that the opportunity to have sex with a relative recurred ancestrally, that it could not be successfully negotiated in the absence of a guilt program specifically taking care of it, that the situation had a rich and reliably repeated structure, that it had reliable cues signaling its presence, and that it was of a type which would have resulted in a large fitness cost (e.g. children who are unhealthy). But the availability of a plausible story about the evolution of incest-caused guilt is clearly not hard evidence for its truth.

As Griffiths (1997) pointed out, adaptationist thinking of this kind can have a "substantive negative heuristic effect", because it not only prevents us from considering alternative non evolutionary explanations, but it also leads us to the wrong expectations concerning the operations of evolution. As he puts it, "[c]ontrary to the predictions of the evolutionary psychologists, affect programs are designed to cope with quite general evolutionary problems, and the affect program system is designed to redefine those problems as the environment changes". These features are apparent in Ekman's research program. He described affect programs as being *open programs* in Ernest Mayr's sense, emphasizing that they evolved "so that we can learn what will work in the particular environment in which we are living" (Ekman 2003, 66). This would be incompatible with having a fear program very specialized in terms of its elicitors, say a lion-caused or even a predator-caused fear program.

Rather, the fear affect program evolved to deal with *danger*, a category at a higher level of abstraction than either predators or lions. This suggests that, even though Ekman regularly quotes Tooby and Cosmides (1990) to illustrate what he means by *fundamental life task*, his understanding of this notion is more general. The expectation that affect programs will deal with general rather than specific evolutionary problems – danger rather than lions – finds further support in the presence of developmental constraints on evolution, which limit the degree to which evolutionary problems can be dealt with independently of one another (Gould and Lewontin 1978).

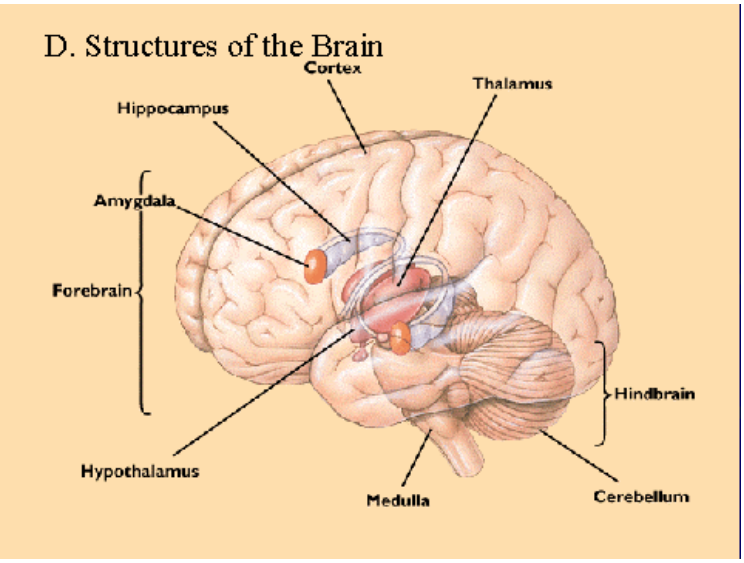
Most importantly, Ekman and other members of the affect program tradition do not accept as evidence for the basic status of an emotion its *mere compatibility* with an adaptationist story. Empirical support is required that a significant number of the 11 characteristics of basic emotions are indeed fulfilled. As a testimony of the stricter criteria of admission to the category of basic emotions, even by the lights of his most liberal list Ekman (1999b) does not think that guilt, love and jealousy, let alone their more specific sub-species, are basic emotions. In other words, he does not think that we have sufficient empirical evidence about their facial signals, physiological changes, presence in other primates, and so on to conclude that the hypothesis of adaptation is well-supported. The evolutionary psychology literature is instead replete with poorly supported accounts of the evolution of guilt, love, jealousy and many other emotions (e.g. Buss 2002; see Buller 2005 for a thorough critical analysis of the evidential support of evolutionary psychology).

What is the evidence offered by affect program theorists, as well as by researchers in other fields, about the evolution of basic emotions? Roughly speaking, there seem to be two main domains of facts supporting the hypothesis that at least some instances of some basic emotions can be given an evolutionary explanation: (a) the nature of emotional appraisal, and (b) the nature of emotional expressions. Let us consider them in turn.

5.2.2. The neurobiology of emotional appraisal

Starting with Zajonc (1980), a flurry of research has investigated the automatic appraisal system associated with emotions, and described its main functional and neurobiological characteristics. It is uncontroversial by now that the appraisal processes characteristic of some tokens of some basic emotion types manifests several of the properties of the input systems

Fodor (1983) called modules (see chapter 10 for discussion). For example, there is strong evidence that the dedicated neural architecture of some forms of emotional appraisal is phylogenetically old, and shared in homologous form across species.



The idea that emotions may not rely on cortical neural pathways finds its first precursors in Cannon (1929) and Bard (1929). The Jamesian account of emotions, the first to address the issue of emotional neurobiology, indicated in the areas of the neocortex that mediate visual, auditory, and somatosensory sensation the brain areas involved in the elicitation of emotions. By means of lesion studies, Cannon (1929) and Bard (1929) brought a first blow to this idea. For example, Bard (1929) showed that the removal of the entire neocortex did not prevent the occurrence of rage responses in animals, whereas the integrity of the hypothalamus, a subcortical region, was required for such responses to be displayed. Cannon and Bard maintained that the *feeling* of emotion depended upon the activation of the neocortex, activated through nerve fibers ascending from the hypothalamus.

Cannon and Bard's intuitions were further elaborated by Papez (1937), who provided the first fairly detailed account of the areas of the brain involved in the elicitation of emotions. What later came to be known as the Papez circuit comprises the hypothalamus, the anterior thalamus, the cingulated gyrus and the hippocampus. A further important paper was published in the same year by Kluver and Bucy (1937), who described a suite of behavioral changes in monkeys brought about by damage to the temporal lobe – a portion of the cerebral cortex. It was reported

that the removal of the temporal lobe in a monkey resulted in the fact that “the animal does not exhibit the reactions generally associated with anger and fear”.

For example, they would no longer respond with fear to the presence of snakes, which previously generated strong emotional reactions. Moreover, the animal lost the ability to recognize what objects are eatable, despite maintaining the ability to visualize them and navigate through them appropriately. Similarly, the animal lost the ability to orient its sexual behaviors to monkeys of the opposite sex, attempting copulation with members of its own sex as well as animals of other species. They described this condition as “psychic blindness” - a syndrome now known as Kluver and Bucy syndrome -, namely the inability to understand the significance of objects in the environment while maintaining the ability to see, smell, touch, taste, and hear them. Weiskrantz (1956) showed that the syndrome could be produced by lesions limited to only one area of the temporal lobe, namely the amygdala.

The turning point in the study of the emotional brain came with MacLean’s (1949, 1952, 1960, 1969) masterful synthesis of several disparate sources of evidence on the role played by the brain in the elicitation of emotions. MacLean (1952) was interested in investigating the neural correlates of emotion, and claimed that we can only understand them if we realize the “hierarchical organization of the brain”. He argued that “man’s brain…has inherited the structure and pattern of organization of three basic types, which, for simplifying reasons, I refer to as reptilian, paleomammalian, and neomammalian [which] must intermesh and function together as a triune brain” (338). The reptilian brain “comprises much of the reticular system, midbrain, and basal ganglia” (338), the paleomammalian brain “is distinguished by a marked outgrowth of primitive cortex, which…is synonymous with the limbic cortex” (338), and the mammalian brain which is distinguished by a neocortex highly differentiated from the primitive cortex, and which is “the hallmark of the brains of higher mammals and which culminates in man to become the brain of reading, writing and arithmetic” (339).

MacLean called limbic system - limbic means “forming a border around” the brainstem - the conjunction of the “limbic cortex and structures in the brainstem with which it has primary connections” (339), substituting limbic system for what he had earlier called *visceral brain* (MacLean, 1949). MacLean included in the limbic system amygdala, septum, hippocampal formation, orbitofrontal cortex and cingulated gyrus. In his impressive synthesis, MacLean marshaled several sources of evidence for the hypothesis that the limbic system is crucially

involved in emotional phenomena. Besides the works by Cannon, Bard, Papez and Kluver and Bucy discussed so far, MacLean reported and discussed a number of further neurobiological studies. For example, MacLean (1970, 341) claimed that "the most convinging evidence that the limbic system is involved in emotional functions" was constituted by the study of human patients with temporal lobe epilepsy.

This kind of epilepsy, as shown by Malamud (1966, 194), is caused by damages of the hippocampal component of the limbic system. MacLean reported that in the proximity of seizures such patients manifested what he labeled *basic affects* (e.g. hunger, thirst, nausea, warmth and the need to defecate or urinate) and *general affects* (e.g. fear, sadness, anger and paranoid feelings). MacLean reported that Penfield et al. (1954) proved that electrical stimulation of the limbic system could produce results analogous to those occurring in cases of naturally occurring seizure. MacLean was also impressed by Hess's (1956) and Hunsperger (1956) studies on the results of the electrical stimulation of cats' brains. Such studies had shown, MacLean (1960) claimed, that angry defensive behaviors in cats – inclusive of facial expressions, postures, specific cardiac activity and action tendencies- , could be produced by direct stimulation of the hypothalamus.

On the basis of these and other sources of evidence, MacLean concluded that the limbic system "is essentially similar throughout the mammalian scale" (1969, 673) and that "in addition to olfactory functions, the limbic cortex is involved in emotional behavior and associated endocrine and viscerosomatic activities" (339). Notably, MacLean also suggested that some of the phenomena Freud tried to capture by positing the existence of the unconscious Id could be explained in terms of the tension between areas of the brain emerged at different points in phylogenetic history.

Since the appearance of MacLean's influential account, there has been much debate concerning the exact role of the limbic system in the generation of emotional phenomena. It has become rather clear that, contrary to MacLean's theory, there isn't a localized system in the brain devoted to the production of emotions. As Le Doux (1996, 99) put it, "MacLean and later enthusiasts of the limbic system have not managed to give us a good way of identifying what parts of the brain actually make up the limbic system". It was shown for example that the distinction between an old and a new cortex was questionable on anatomical grounds (Le Doux 1996, 100). Also, the idea that higher cognitive functions are served by the neocortex exclusively

was put into question by studies that show that declarative memory can be impaired by damaging the hippocampus in a way that preserves instead the ability to produce emotional responses. There are two possible responses to the mounting evidence that the limbic system as understood by Maclean has at best unclear identity conditions.

One is championed by Le Doux (1996) and Brothers (1997), who argued that "the limbic system term…is imprecise and…should be discarded" (Le Doux 1996, 101). The other is championed by Panskepp (2000), who suggests that we should keep using the term as long as we understand it as "higher-order conceptual entity that helps us designate and discuss the general locations of the families of functional neural systems commonly placed under the conceptual umbrella "emotion"" (139). Panksepp's (2000) idea is to employ the idea of limbic system to designate and study "the general brain areas that are especially influential in elaborating emotions" (140), thereby defining the limbic system not in terms of its anatomy and phylogenetic history but, rather, in terms of its function. Panskepp (2000) ultimately thinks that "credible answers to out affective questions will gradually come as we unravel the integrated brain structures that mediate the best external signs of emotionality we can agree upon" (139).

I will not take position on this controversy here. What I take to be important about the neurobiological evidence I described are two things in particular. Firstly, that it offers strong evidence for the evolutionary continuity of human and animal basic emotions. Secondly, that it provides a neurobiological mechanism for making sense of an idea we have seen hinted at over and over again in the history of emotion theory, namely that emotional appraisal is in many cases primitive and does not recruit complex cognitive abilities. Even if we discard the idea of an all-purpose system for the emotions, the fact that phylogenetically old circuits of the brain may mediate some forms of emotionality maintains its import. Le Doux (1996) is both a critic of the limbic system hypothesis, and the main contributor to the elucidation of the neural underpinnings of fear, the best understood emotion from the neurobiological point of view.

LeDoux (1996) demonstrated by means of ingenious lesion studies that there is a kind of fear elicited in a reflex-like fashion through a neural *low road* that bypasses the neocortex, and projects along a subcortical pathway directly to the amygdala. This form of fear is supplemented by *high road* fear, which projects instead to the amygdala indirectly through the sensory cortex. The picture below summarizes the two neural pathways to fear as described by LeDoux (1996):

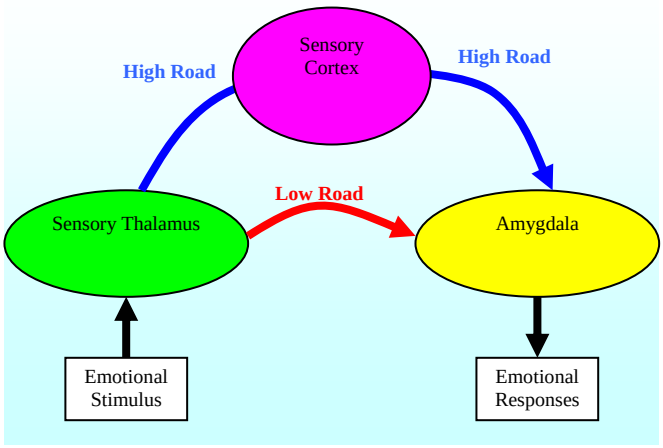


Figure 4: Le Doux's high and low pathways to fear

LeDoux (1996) persuasively argued that the neurobiological discovery of a low road to fear indicates that "emotional responses can occur without the involvement of the higher processing systems of the brain, systems believed to be involved in thinking, reasoning, and consciousness" (1996, 161). The fact that fear conditioning demonstrably occurs along subcortical neural pathways provides evidence to the conclusion that there is a kind of fear which can be found in homologous form in distinct species.

A further strand of research on fear appraisal offers evidence that fear was selected for by natural selection. Seligman (1971) reasoned that, if fear is an adaptation, then appraisal should be especially sensitive to things which have been dangerous in ancestral time, such as spiders or snakes or threatening expressions. Recently, Öhman (1999, 2002) has applied the backward masking technique developed by Marcel (1983) to argue that we are in fact prepared to fear spiders, snakes and threatening expressions more than other things. Backward masking consists of the presentation of two stimuli in rapid succession, a *target stimulus* and a *masking stimulus*. When the interval between the two presentations – the time of stimulus-onset asynchrony (SOA) - is sufficiently short (30 ms or less), the masking stimulus effectively *masks* the target stimulus.

In response to questions, subjects report that they are not aware of being exposed to the masked visual stimulus, in the sense that they are unable to verbally report on it. Moreover, when forced to guess on the characteristics of the masked stimulus, they perform at chance level.

Ohman demonstrated that fear conditioning to a non-masked stimulus is quicker when the stimulus is a snake or a spider or a threatening face than when it is a currently dangerous object (e.g. a gun) or a neutral one (e.g. a flower). Also, he demonstrated that whereas the fear response disappears with masked guns and flowers, it is maintained towards masked images of spiders, snakes and threatening faces, suggesting that awareness is not a requirement for the elicitation of fear responses to ancestral challenges. This suggests that fear of spiders is an adaptation, selected for in the ancestral past when it served some useful purpose.

5.2.3. Facial expressions and evolution

Ekman's main scientific contribution consists of having collected a large amount of evidence in support of the thesis that basic emotions have distinctive universal signals. To this achievement, Ekman and his collaborators have added some preliminary evidence that at least some of the basic emotions have different patterns of autonomic nervous system activity. I begin from the latter contribution, which can be dealt with more quickly. Ekman predicted that physiological responses should be distinctive for each basic emotions, and invariant across cultures. This goes against Cannon's (1929) critique to the feeling theory, which I reported in chapter 3. But although Ekman, Levenson & Friesen (1983) and Levenson, Ekman & Friesen, (1990) have presented some evidence that anger, fear, disgust and sadness have different autonomic signatures, many other studies have offered contrary evidence (e.g. Ax 1953, Malmo 1950). At this stage of research, Ekman (1999b, 49) himself acknowledges that "the matter is far from being completely settled".

I want to argue that we should not expect that if emotions are adaptations then they should have a different autonomic signature. Ekman posited this hypothesis because he thought that, if emotions evolved to deal with fundamental life tasks, then the physiology of the body should reflect optimal preparation to adaptive action when faced with them. From this it does not follow, however, that physiological changes must be differentiated (see my discussion in chapter 10). As Ekman (1999b, 50) himself remarks, if "no specific pattern of motor activity had survival value for an emotion, then there would be no reason to expect a specific pattern of ANS activity to have been established for that emotion". I want to add to this point that there may in principle be life tasks such that dealing with them adaptively does not even require autonomic activation, let alone a distinctive one.

What I find quite intriguing is a preliminary finding that the same autonomic activity is found in different cultures, because in this case the hypothesis that physiological responses results from species-constant learning is not very persuasive. Levenson et al. (1992) offered some evidence that the physiological profile of some tokens of some basic emotions is invariant in Western and Non-Western cultures. They showed that the Minangkabau of Western Sumatra have the same emotion-specific physiology detected by Ekman, Levenson & Friesen's (1983) with respect to anger, fear, disgust and sadness

Jointly with the neurobiological evidence I discussed in the previous section, the main source of evidence for the thesis that emotions are adaptations comes from the study of emotional expressions. In particular, there is evidence that (a) some emotional expressions are present at birth or emerge in the first few months of life, and they express the same emotions in infants as they do in human adults (b) some emotional expressions are present in homologous form in species related to man, and they express the same emotions in animals as they do in human adults (c) some emotional expressions are present in all cultures, and they express the same emotions in each of them. In 1978, Ekman and Friesen developed the Facial Action Coding System (FACS), a system to measure the pattern and timing of facial movement in an objective and precise manner (updated in Ekman et al. 2002). The system was extended in 1992 to infants, with the creation of the Baby FACS by Oster and Rosenstein (1992).

Let us consider a few highlights from the literature on emotional expressions in infants and animals. When they are as young as 3 weeks old, infants produce the facial expression of joy when exposed to a friendly human face, and at 8 weeks they produce the facial response of anger when in pain (Izard 1994). As summarized by Izard, "[b]y 2.5 months, infants encoded full-face and partial expressions of interest, joy, sadness, and anger with sufficient frequency of statistical analysis" as well as "surprise, disgust, and fear", although the latter three not often enough for a statistical analysis. Moreover, Izard demonstrated that these emotional expressions in infants recruited the same muscles they recruit in adults.

Using a version of the Baby FACS, Camras (1992) proved that arm restraint generates the same facial expressions of anger in 5 months old American and Japanese infants, and that these expressions are morphologically identical. Studies of children born blind also confirm that they manifest the same facial expressions of emotion as non-blind children (Goodenough, 1931, Thompson, 1941, Eibl-Eibesfeld 1973). Concerning facial expressions in related species, the

primatologists Van Hooff (1967), Chevalier-Skolnikoff (1973) and Redican (1982) reviewed the literature on facial expressions in New and Old World monkeys, and agreed that humans and monkeys show a number of the same expressions. For example, Chevalier-Skolnikoff (1973) argued that there is homology between the expressions of human and simian anger, sadness, and affection, as well as in the crying and laughter expressions. As I reported, Darwin considered the similarities in facial expression between humans and related species to be an important source of evidence for the theory of evolution.

Most of the evidence for the evolution of facial expressions comes from cross-cultural studies, which have tried to establish that some emotions are expressed in the same way in all cultures. I now turn to the debate generated by such studies, which will lead us to the emergence of the social constructionist paradigm.

5.2.4. Critiques of Darwin's universality thesis

For 90 years, *The Expression of Emotions in Humans and Animals* went mostly unrecognized. Part of the explanation for this neglect is that the universality thesis about emotional expressions, namely the thesis that some emotions exist and are expressed in the same way in all cultures, was considered to be flawed. Three influential sources of evidence were presented in the 1920s, 1930s, and 1940s in favor of the view that emotional expressions are not universal and more generally not inherited.[7] Firstly, Floyd Allport (1924) introduced a theory of species-constant learning, according to which certain traits which appear invariantly in different cultures are learned in all cultures because of their usefulness, rather than inherited from a common ancestor. This means that, even if the same expressions can be proven to appear in every culture, the hypothesis that they are inherited is questionable.

Secondly, Landis (1924) denied both that emotions have specific emotional expressions associated to them, and that such expressions could be recognized by observers. He took pictures of 25 subjects in 17 emotion-inducing contexts, and failed to detect a consistent emotion-expression relationship. In 1929, Landis selected some of the most expressive pictures from his first study, and showed them to observers, who allegedly failed to recognize them.

Thirdly, Klineberg (1940) questioned the assumption that similar emotions could be found in species related to men. He based this conclusion on the evidence that human beings could not

[7] This reconstruction is heavily drawn from Russell (1994)

recognize the facial expressions of chimpanzees, as reported by Foley (1935). His view was that "the great difficulty experienced by untrained human observers in recognizing the emotions of chimpanzees from their facial expressions strengthens the hypothesis of cultural or social determination of the expressions of emotions in man". Klineberg's (1940) conviction was that "[e]motional expression is analogous to language in that it functions as a means of communication, and that it must be learned, at least in part" (179, 200).

Klineberg detailed various examples of cross-cultural differences in the expression of emotion, for example the protrusion of the tongue in surprise reported in a Chinese novel. However, he also pointed out that "undoubtedly certain types of expressive behavior … are common to all human societies" (Klineberg, 1940, 176). Among the examples, he listed laughing, crying and trembling. Even with respect to such expressions, however, Klineberg suggested the presence of cultural rules. For example, even if crying is universally present in all cultures, there are cross-cultural differences concerning how and when one should cry. He gave the example of the intensity of weeping in grief, which manifests cross-cultural variation.

Bateson and Mead (1942, 39) reported on cross-cultural differences in emotional instrumental behaviors. It was argued that the Balinese express fear by falling asleep. As they stated, "[t]he child who is frightened by the tantrum of his child nurse falls asleep as she shrieks out her unrestrained rage right beside his closed ear. The older child who has lost or broken some valuable thing will be found when his parents return…in a deep sleep…Children learn to be afraid of birth, and if they find themselves in the house … with a birth, they fall into a deep sleep…. The thief whose case is being tried falls asleep". By the 1950s and 1960s, the idea that emotional expressions are universal had largely been put to rest. In her 1955 introduction to Darwin's book on emotional expressions, Margaret Mead stated that the publisher had had the felicitous idea of "adding at the end of the book some examples of recent work which carry on the inquiry which Charles Darwin initiated" (Darwin 1955, vi). The list of such works clearly indicates that by then some of the most influential writers on expression had become supporters of the anti-universality thesis.

She did mention among the additions "Konrad Lorenz's drawings on expressive behaviors in animals", but also "selections from Gregory Bateson's photographic studies on the Balinese", and photos of members of "the group taking part in the new science of kinesics". The last remark is especially telling, because in the 1960s and 1970s Ray Birdwhistell's science of kinesics – the

science of movements and facial expressions - , had become the most influential source of skepticism about the universality thesis. Birdwhistell (1970) wrote:

> When I first became interested in studying body motion I was confident that it would be possible to isolate a series of expressions, postures and movements that 'very denotative of primary emotional states... As research proceeded, and even before the development of kinesics, it became clear that this search for universals was culture-bound... There are probably no universal symbols of emotional state. ... We can expect them [emotional expressions] to be learned and patterned according to the particular structures of particular societies (126).

As evidence for this thesis, Birdwhistell reported that fact that in his studies of the human smile he had found over and over again that subjects smile in both favorable and aversive conditions. What can we make of the critiques we have so far collected? Let us consider them one by one, starting from Allport's (1924) hypothesis that emotional expressions could result from species-constant learning. What goes against it, as argued by Griffiths (1997), is the fact that the emotional expressions we find cross-culturally are *arbitrarily* associated to the emotions they express. There does not appear to be a specific usefulness in expressing, say, anger with narrowed rather than open eyes, or fear with hair erect rather than not erect. This being the case, if usefulness were indeed the only reason why emotions are constantly learned in different cultures, we should find a proliferation of different expressions for the same emotions in different cultures. The fact that we do not find it suggests that the universality of expressions must have an explanation other than the fact that every culture finds it useful to express emotions and learns how to do that from scratch.

Concerning Landis' (1924) thesis that observers disagree on what emotion is expressed by a certain picture, a great amount of evidence has been offered since then to prove it false. Ekman (1972) criticized the specific set-up of Landis' (1924) study, and, jointly with Izard and other researchers, he offered a wealth of evidence to the effect that at least some emotions are recognized cross-culturally. The most common cross-cultural technique used by Ekman and his associates is a version of Darwin's own picture technique. It consists of showing pictures of emotional expressions and asking observers what emotions they express from a list of six to ten emotion terms in the observer's language. The following are examples of the expressions used:

Figure 5: Facial expressions of happiness, surprise, fear, anger, disgust, sadness

As reported by Ekman (1999a), these experiments have so far been performed with observers from 21 literate countries: Africa, Argentina, Brazil, Chile, China, England, Estonia, Ethiopia, France, Germany, Greece, Italy, Japan, Kirghizistan, Malaysia, Scotland, Sweden, Indonesia (Sumatra), Switzerland, Turkey and the USA (Ekman, Sorenson & Friesen 1969b, Izard, 1971; Niit & Valsiner, 1991; Boucher & Carlson, 1980; Ducci, Arcuri, Georgis, & Sineshaw, 1982; McAndrew, 1986, Ekman et al., 1987). In all these experiments, pictures of happiness, anger, fear, sadness, disgust and surprise were used, plus other expressions specific to the particular study. Here is Ekman's (1999a) summary of the evidence:

> There was an extraordinary amount of agreement about which emotion was shown in which photographs across the 21 countries. In *every* case, the majority in each of the 21 countries agreed about the pictures that showed happiness, those that showed sadness and those that showed disgust. For surprise expressions there was agreement by the majority in 20 out of the 21 countries, for fear on 19 out of 21, and for anger in 18 out of 21. In those 6 cases in

> which the *majority* did not choose the same emotion as was chosen in every other country, the *most frequent* response (although it was not the majority), was the same as was given by the majority in the other countries. In my own studies, the only studies in which the expressions were selected on the basis of measuring the muscle movements shown in the photographs, *all* the expressions were judged as showing the same emotion by the majority in *every* country we studied (305, 306).

These results, criticized by Russell (1994) because of the forced-choice schema, have been to some extent replicated when subjects were asked to associated an emotion term to the expression freely rather than from a list of pre-established choices (e.g. Izard 1971, Boucher & Carlson 1980, Rosenberg & Ekman 1993). A problem with the studies cited so far is that they only deal with literate cultures. In principle, such cultures may not really be isolated from one another, in the sense that by accessing the same visual representations of emotions (e.g. in movies) people may simply learn what some expressions stand for, and either import them in their own culture or simply become able to recognize them. Under this view, the presence of an arbitrary universal expression across cultures would still be compatible with the hypothesis of species constant-learning.

In 1967, Ekman went to study the South Fore culture in Papua New Guinea, a culture that has no access to photographs, movies, or magazines of any kind and no written language. Ekman and Friesen (1971) demonstrated that the natives were able to associate a story designed to elicit a particular emotion, say sadness, with the facial expression of sadness. This turned out to be the case with respect to the six emotions of happiness, anger, fear, sadness, disgust and surprise, although the natives were not able to distinguish the expression of surprise from that of fear. Russell (1994) criticized such results on the basis of a report by Sorenson (1976), who traveled with Ekman and Friesen and argued that “it was likely that at least some responses were influenced by feedback between translator and subject”, concluding that the possible “leaking” of cues concerning the desired answer “undoubtedly skewed our results” (Sorenson, 1976, 139–140).

Ekman (1994) responded to have taken all possible precautions to avoid leaking, and dismissed the criticism, although admitting that some form of influence might have been possible in principle. Ekman (1972) also asked South Fore people to generate the emotional expressions associated with each story, obtaining expressions such as the following:

Figure 6: Facial expressions from Papua New Guinea

When he presented these expressions to American students, he found out that they were able to associate to the New Guinean expressions the emotion the story was meant to elicit, once again with the exception of surprise and fear, which were not distinguished from one another.

The third strand of criticism of the universality thesis was that there are emotional expressions typical of some cultures but not of others. I discussed Klineberg's report on the differences in intensity of weeping across cultures, Bateson and Mead's (1942) report that fear is associated to sleeping among the Balinese, and Birdwhistell's claim that smiles are expressed in both favorable and aversive conditions. What could one say of such cases? Since the beginning, Ekman's position was not to deny the possibility of emotional expressions (and other characteristics of emotions) unique to some cultures, but rather to affirm the presence of emotional expressions common to all cultures. The fact that the Balinese have a culturally-specific behavior associated with fear does not prove that there is no universal fear expression. Birdwhistell's concern can be answered in light of Ekman's (1980, 79) remark that expressions such as "frown, smile, play-face and even brow-raise are much too gross" to do science with.

Once expressions are carefully distinguished through the FACS system, it becomes clear that there are at least two kinds of smiles, namely the smile that expresses enjoyment (with lip

corners pulled up and contracted muscles around the eyes) and smiles which express things other than enjoyment (without lip corners pulled up and without contracted muscles around the eyes). Since Duchenne was the first to notice this difference, Ekman (1992) called the smile associated with enjoyment *Duchenne smile*, and distinguished between several varieties of non Duchenne smiles.

In light of these distinctions, Birdwhistell's report that people smile in unfavorable circumstances is no longer problematic, as long as the smile is not a Duchenne smile. Ekman's claim is only that the Duchenne smile is a universal expression of enjoyment. Klineberg's point that even universal expressions such as weeping are different in different cultures was developed by Ekman into the notion of a *display rule*, which makes the universality thesis compatible with cultural variation. Ekman wrote: "While the facial muscles which move when a particular affect is aroused are the same across cultures, the evoking stimuli, the linked affects, the display rules and the behavioral consequences all can vary from one culture to another" (Ekman and Friesen, 1969, 73).

Since the beginning, he called his approach to facial expressions "neurocultural", in order to emphasize "two very different sets of determinants of facial expressions, one which is responsible for universals and the other for cultural differences" (Ekman 1972, 212). In 1973, Ekman offered a general account of *display rules*, distinguishing between emphasizing, de-emphasizing, dissimulating and simulating rules. Evidence for the presence of such rules was offered when Ekman (1972) studied the expressions of American and Japanese students exposed to stressful movies while their faces were secretly video-recorded. The morphology of facial movements turned out to be very similar when the students watched the movie alone, compatibly with the universality thesis. However, when a white-coated research assistant entered the video room while the students were watching the video, the Japanese students, differently from their American counterparts, started dissimulating their negative emotions through smiling.

Ekman argued that the culturally specific display rule did not prevent the automatic negative expression from being present in a subdued form, as a micromovement analysis of the video recording reveals. The experiment was criticized by Fridlund and Duchaine (1996), who argued that Japanese and American students may have shown different facial expressions in the presence of an authority figure because they experienced different emotions, not because they followed different display rules for the expression of the same emotion.

Despite the limitations of the specific experiments I discussed, if we consider the literature on facial expressions in infants, animals and different human cultures as a whole, it seems clear that they support the hypothesis that some emotions are adaptations. This is especially true with respect to the emotions of happiness, anger, disgust, sadness and fear/surprise. Some preliminary evidence has also been marshaled in favor of the hypothesis that contempt is universally expressed (Ekman & Friesen, 1986; Ekman & Heider 1988; Matsumoto, 1992). Keltner (1995) has offered some evidence for the universal facial expression of embarrassment. On the other hand, so far no significant evidence has been offered for the universality of the expressions of excitement, guilt, pride in achievement, relief, sensory pleasure, and shame, as well as for the universality of the expression of the emotions which do not meet Ekman's requirements for being basic (e.g. hatred, romantic love, grief, jealousy, interest).

5.3. CONCLUSION

I have explored the evidence for the thesis that basic emotions are adaptations. The most compelling sources of evidence have to do with the neurobiology of emotional appraisal and with the nature of emotional expressions. So it looks like there are at least some basic emotions, as affect program theorists understand them. Even though affect program theorists claim that there are no emotions other than basic ones, it is not clear what the rationale for this claim would be. As I shall argue in chapter 9, the problem is once again that emotion theorists lack a clear understanding of the kind of activity in which they are engaged when they try to offer an account of what the emotions are.

6. EMOTIONS AS SOCIAL CONSTRUCTIONS

As the evolutionary tradition, the social constructionist tradition is interested in figuring out what emotions are from studying what sorts of problems they solve. Social constructionists, however, focus mostly on the social functions played by the emotions. They are convinced that, if there is anything biological about the emotions, it is not central. I think the emphasis social constructionists have put on the social dimension of emotions is a welcome one. Whatever we think about the specific proposals emerging from this tradition, one thing is certain, namely that they have raised our consciousness about what has been neglected for a long time, namely the impact emotions have on interpersonal relations.

The social constructionist approach found its first proponents in the 1920s, when a number of anthropologists and social scientists, as I described in the previous chapter, started questioning Darwin's (1872) evidence for the universal nature of emotional expressions. Sartre can be considered one of the first to offer a general, although idiosyncratic, theory of emotions as social roles, a view developed in the early 1980s by philosophers (e.g. Harre 1986, Armon-Jones 1985, 1986a, 1986b), psychologists (e.g. Averill 1980), and anthropologists (e.g. Lutz 1988).

In recent times, Parkinson (1995, 1999, 2004) and especially Griffiths (2003, 2004a) have tried to reconcile the evolutionary and the social constructionist traditions, taking their cue from some important work on emotional expressions by Fridlund (1989, 1997), Russell (1997) and Fernandez-Dols (1997a, 1997b) among others (see Russell et al. 2003 for a review). The basic insight of what we may call the *strategic socio-evolutionary* strand of the social constructionist tradition is that the communicative side of the emotions has fundamentally strategic dimensions, which contribute to explain both why the emotion evolved and why they must be understood in the context of a social transaction.

6.1. WHAT IS SOCIAL CONSTRUCTIONISM?

In the early 1980s, a wave of theorizing about the emotions appeared under the general label of *social constructionism*. Contributions commonly gathered under this label span a number of disciplines, including philosophy (Coulter 1986, Harre 1986, Armon-Jones 1986), social theory (Sabini and Silver 1982), psychology (Averill 1980), and anthropology (Lutz 1982). One of the great problems of social constructionist theses in general is that it is not clear what is ascribed to an entity when it is argued that it is socially constructed. Our first order of business is therefore to get clear on the very idea of a *social construction.* My account of this issue relies on Hacking's (1999) discussion, which offers a general framework I will then apply to the emotions.

Hacking (1999) noticed that social constructionist claims are currently quite fashionable in a number of contemporary disciplines, and that they are made, rather nonchalantly, with respect to a large variety of entities (e.g. authorship in Woodmansee and Jaszi 1994, the child viewer of television in Luke 1990, danger in McCormick 1995, facts in Latour and Woolgar 1979, gender in Dewar 1986, quarks in Pickering 1984, women refugees in Moussa 1992, etc.). Generally speaking, the point of social constructionist claims about some X is for Hacking "to raise consciousness" (6) about the following thesis:

> (1) X need not have existed, or need not be at all as it is. X, or X as it is at present, is not determined by the nature of things; it is not inevitable (Hacking 1999, 6)

Call this the *social contingency thesis*. To say that X need not have existed as it is, or that it is not inevitable, is to say, as suggested by Kukla (2000, 2), "that not-X is possible". But there are various notions of possibility, and not all of them capture what social constructionists want to raise consciousness about. The key claim is that X in its present form is not "determined by the nature of things", in the sense that X might have been or be different if society/social arrangements/social circumstances/social interaction/social selves/social values/social forces/social events/social facts/social history/social needs/social interests had been or were

different in a way that is still compatible with the nature of things. This is the subjunctive conditional on the alleged truth of which social constructionists want to call our attention.

For example, one may say that X=marriage would not be the way it is in the US had the US embraced the values of an Islamic society. Or one may say that if blacks were the majority in the US, X=black race would not be the way it is. These claims appear to be platitudinous, which brings us to a requirement that social contructionist theses must fulfill in order to avoid triviality. Hacking refers to the following as a "precondition" for social constructionist ascriptions:

> In the present state of affairs, X is taken for granted; X appears to be inevitable (Hacking 1999, 12)

The reason why we "do not find books on the social construction of banks, the fiscal system, checks, money, dollar bills, bills of landing, contracts, tort, the Federal Reserve, or the British Monarchy" (12), tells us Hacking, is precisely that (0) is not fulfilled, namely that the truth of the social contingency thesis about X is obvious to everyone. The necessity of precondition (0) can be used as a heuristic to find out what social constructionists mean by their claims. If X under some interpretation is *obviously* dependent on contingent social factors, then maybe it is not under that interpretation of X that the claim is made.

Suppose a social constructionist were to say that "women refugees are socially constructed". Interpreted as a claim about members of the extension of the kind *woman refugee*, the assertion is trivially true, as women become refugees by virtue of contingent social circumstances. Hacking sensibly suggests that the right way to interpret the assertion is in terms of a claim about the idea of *woman refugee*. As Hacking (1999, 11) puts it, "when we read of the social construction of X, it is very commonly the idea of X . . . that is meant". If pre-condition (0) has been fulfilled, calling attention on the truth of the social contingency thesis about X does not amount to stating the obvious. However, there still is the need of a motivation for raising consciousness about the truth of a non-trivial thesis.

Hacking suggests that social constructivists "very often … go further [than thesis (1)], and urge that" (6):

> (2) X is quite bad as it is

> (3) We would be much better off if X were done away with, or at least radically transformed (Hacking 1999, 6)

Hacking qualifies his position by adding that "[m]any social construction theses at once advance to (2) and (3), but they need not do so" (7). They need not do so because it may very well be the case that a socially contingent X is not bad, or that changing X would not be worth the cost involved in doing so, and so on. But if theses (2) and (3) are not advanced, the question is still open of why anyone would want to raise consciousness about the truth of (1). Hacking makes a number of distinctions between varieties of social constructionism in terms of their basic motivation, but such distinctions are not relevant for my purposes.

I distill from Hacking's account two general and not mutually exclusive *motivations* that may ground the issuing of a social constructionist ascription. One is the *intellectual motivation* of furthering our understanding of X. The idea here is that if we really want to understand X and the phenomena in which X is involved, we must understand how X is crucially dependent on contingent social aspects. The other is the *political motivation* of mobilizing resources to change X. The idea here is that we ought to change X and the phenomena in which X is involved, and once we understand that there is nothing inevitable about it we will be in a better position to get rid of it or modify it to some extent. To conclude, social constructionists aim to raise consciousness, for intellectual and/or political reasons, about the fact that it is not the *nature of things*, but rather our *contingent social arrangements*, that make X what it is. On the basis of this preliminary account of social constructionism, we can turn to an understanding of social constructionism about the emotions.

6.1.1. Two strands of social constructionism about emotions

Social constructionism about the emotions can be understood as a *set of interpretations* of the *social contingency thesis* that "Emotion/anger need not have existed, or need not be at all as it is. Emotion/anger, or emotion/anger as it is at present, is not determined by the nature of things; it is not inevitable". The motivation for raising consciousness about the social contingency of emotions is primarily *intellectual*. What social constructionists want to tell us is that if we want to *understand* emotions and the phenomena in which they are involved, we need to pay attention to the fact that, in various senses we need to understand, emotions are not

"determined by the nature of things". As Averill (1980) puts it, "[t]he emotions can only be fully understood as part of the culture as a whole" (315).

We find occasional hints of a more *political* nature, for example when Armond-Jones (1986a) tells us that since "emotions are functionally constituted for the maintenance of particular value systems, then a moral issue arises in that such organizations of human experience not only are socially based, but also can be evaluated as desirable and just" (35). By and large, however, social constructionism about the emotions is presented with the intent of raising consciousness for the sake of *understanding*, rather than for the sake of mobilizing resources for *political change*.

Social constructionists aim to raise the emotion theorist's consciousness about the fact that emotions are what they are because of society/social arrangements/social circumstances/social interaction/social selves/social values/social forces/social events/social facts/social history/social needs/social interests, rather than because of nature. The way in which this position must be interpreted, I suggest, is in terms of *emphasis* on the social rather than on the natural. It seems to me that social constructionists are not saying that nothing about the emotions is natural, a position that would clearly be absurd. Rather, what they are saying is that theorists of the emotions have so far put too much emphasis on the natural, disregarding social aspects of the emotions which are essential to understanding what they are. The reversal of emphasis they urge is well-portrayed in the following passages from the social constructionist literature:

> Historically, there has been a tendency to treat emotions as biologically primitive, instinctive response patterns. It is against this backdrop that the concept of emotions as social constructions must be viewed. It has, of course, long been recognized that emotional expression and eliciting conditions are subject to cultural influence. Nevertheless, theorists have tended to treat cultural differences in emotion as superficial variations imposed on basic biological substrata…That is, there is some "core" aspect of emotional behavior, identifiable in terms of neurobiological circuits and/or subjective experience, which is biologically given and hence pan-cultural. By contrast, a basic assumption of the present chapter is that there are no core aspects of emotion which are not intrinsically and essentially influenced by sociocultural factors (Averill 1980, 58).
>
> [My aim is] to deconstruct an overly naturalized and rigidly bounded concept of emotion, to treat emotion as an ideological

> practice rather than a thing to be discovered or an essence to be distilled (Lutz 1988, 4)
>
> Emotions are not just remnants of our phylogenetic past, nor can they be explained in strictly physiological terms. Rather, they are social constructions, and they can be fully understood only on a social level of analysis (Parkinson 1995, 309)

In these statements, we see the natural/social dichotomy embodied by contrasts such as the one between the *social* and the *naturalized*, the *social* and the *strictly physiological*, the *social* and the *biologically primitive*. The dichotomy takes a number of other shapes in the social constructionist literature on the emotions. For example, the social is sometimes contrasted with the *innate*, the *genetically programmed*, the *not-learned*, the *inherited*, the *evolutionarily adaptive*, and the *bodily*.

I am very suspicious of these dichotomies, because I endorse an interactionist view of phenotypic traits such as the one proposed by Kitcher (2001) or Schaffner (1998). Under this view, all phenotypic traits result from an interaction between nature and nurture, and the distinction between the contribution of biology and the contribution of culture is in most cases moot (although not in all cases). But I do not think social constructionism should be judged in terms of how well founded the dichotomy between the natural and the social is. Rather, it should be judged in terms of whether or not the features of the emotions social constructionists consider to have been neglected in emotion theory can really contribute to our understanding of what the emotions are.

The fact that social constructionists consistently call such features *social*, and contrast them with *natural* ones, is to be considered as nothing more than a rethorical consciousness-raising tool. I consider social constructionism about the emotions to have been motivated by two main theses:

(a) Emotions are different in several essential respects in different cultures

(b) Emotions fulfill ends by virtue of which they should be considered actions or roles or moves rather than passions

Call the first the thesis of *emotions as culturally specific syndromes*, and the second the thesis of *emotions as roles/transactions*. Let us consider them in turn.

6.2. EMOTIONS AS CULTURALLY SPECIFIC SYNDROMES

The evidence on cultural variation in the emotions' domain can be organized under two headings: (a) variation in the marks of emotionality, (b) variation in the structure of lexical emotion categories

6.2.1. Do emotions differ in different cultures?[8]

I argued that prototypical instances of prototypical emotions tend to have a variety of marks, including appraisal, physiological responses, expressions, instrumental behaviors and mental actions. I pointed out that starting in the 1920s the hypothesis that emotional expressions are universal had been questioned on the basis of evidence that there are culturally-specific expressions. The same cultural specificity has been demonstrated with respect to all other marks of emotionality. It has been shown for example that the same emotions are elicited by different antecedents in different cultures, which means that the way people appraise emotion-causing events is to some extent culturally specific. Among other things, this will have an impact on the frequency of occurrence of a certain emotion in a certain culture.

Differences in what elicits what emotion may simply be due to different material conditions in the cultures compared (e.g. wealth, climate, population density, risk factors, etc.). For example, in a culture in which restaurants are blown up often, eating in a restaurant will be appraised as dangerous and elicit fear, whereas in a peaceful culture eating in a restaurant will not be appraised as dangerous. Also, every culture will have many elicitors of emotions not present in other cultures (e.g. specific animals for fear, specific forms of entertainment for joy, etc.).

More interestingly, differences in appraisal may result from differences in the views of the self, in the "focal concerns" (Frijda and Mesquita 1992), and in the norms that govern a given culture. Markus and Kitayama (1991) have argued for example that Westeners tend to focus on their individual achievements and autonomy, whereas in several non-Western cultures people are more focused on relationships and interdependence. This is likely to have an impact on what elicits certain emotions, and on how frequently they occur. Markus and Kitayama (1991) have

[8] This section is a close summary of Mesquita and Frijda (1992), one of the best available reviews of cultural variation in the emotions. Most of the references I cite are lifted directly from the bibliography of their article.

proposed that this difference explains why events affecting relationships are frequent elicitors of emotions in Japan, a highly interdependent culture. This results in the fact that, as they argue, emotions such as shame and respect are more prominent in Japan than in the US, whereas emotions such as anger and pride are more prominent in the US than in Japan.

A given culture may have a specific value system which affects what kinds of things elicit what emotions. For example, among the Awlad'Ali, a Beduin tribe living in Egypt, one the most prominent shared concerns – a focal concern - is that of honor. This turns into elicitors of shame many events we would not consider to be shameful, such as the simple interaction with members of a more powerful tribe (Abu-Lughod 1986). Notably, the Awlad'Ali sometimes appraise the death of a loved one as an anger elicitor, because they consider public sadness to be a threat to their personal honor (Abu-Lughod, 1986). In cultures in which the focal concern is that of societal communion, just being alone can become en elicitor of sadness (Briggs 1970).

Japanese people have been reported to appraise injustice as an anger elicitor much more rarely than Europeans and Americans (Scherer et al. 1988). On the other hand, Japanese people blame themselves more easily for negative events, even when they are not directly responsible for them. For example, the event of being cheated on by one's husband often becomes a guilt elicitor among Japanese women (Lebra 1983), whereas it would most likely elicit anger among Americans and Europeans. Sometimes the nature of appraisal is affected by norms about the value of the emotion itself. For example, Briggs (1970) argues that anger is widely disapproved among the Utku Eskimos, who consider it highly disruptive of the social order. As a consequence, very few events are appraised as worthy of anger, which is mainly directed towards dogs (Briggs 1970). On the other hand, in the Kaluli tribe anger is often considered to entitle one to compensation, and many events are appraised as anger elicitors (Schieffelin 1983, 186).

Physiological reactions across cultures have not been studied much (but I reported on Levenson et al. 1992's preliminary evidence in support of universality). Cross-cultural studies have instead been made concerning self-reports of the physiological underpinnings of emotions. Scherer et al. (1986, 1988) asked students from different European countries, Israel, United States and Japan to describe the bodily reactions correspondent to anger, happiness, sadness, joy, and fear, and they obtained very similar results for the first three groups (e.g. unpleasant arousal, lowering of temperature, blood pressure increase, gastric sensations, sweating, and muscular tension for fear). Japanese subjects, on the other hand, reported less physiological sensations that

European, Israeli and American subjects. Low focus on bodily symptoms was also reported in spontaneous description of emotions among the Samoans (Gerber 1985) and the Ifaluk (Llutz 1987). Using a forced choice questionnaire which included lists of physical symptoms (e.g. muscles tensed, heartbeat, etc.), however, Wallbott and Scherer (1988) were able to report strong similarities between physiological self-ascriptions in 27 countries. These data do not answer one way or the other whether or not there are actual physiological differences in the experience of emotions, as they only address what bodily symptoms people typically associate with the emotion categories they use.

Concerning emotional expressions, I have already discussed the topic in the previous chapter. I pointed out that even affect program theorists acknowledge the presence of *display rules* which shape emotional expressions (e.g. the rule of suppression of negative emotions when in the presence of authority manifested by the Japanese students in Ekman 1972), as well as expressions specific to a culture. An example of the latter is given by Shweder (1991, 246), who reports an expression common among the Oryia women in a region of India "in which the tongue extends out and downward and is bitten between the teeth, the eyebrows rise, and the eyes widen, bulge, and cross", associated to a sort of surprise/embarrassment/fear. Cultural differences have also been reported with respect to the instrumental behaviors characteristic of emotions. We have already discussed the case of the Balinese, who reportedly fall asleep when afraid (Bateson and Mead 1942).

Different views of the self, different focal concerns, and different norms are likely to influence what behaviors are associated to what emotions. For example, in a culture such as the Kaluli one, which approves of anger and even compensates for it, "[w]hen a man has suffered wrong or loss…, he may stamp furiously up and down the outside yard or inside hall of the longhouse yelling the particulars of his injury for everyone to hear" (Schieffelin, 1983, 186). In cultures which disapprove of anger, instead, the behaviors associated with it are likely to be much more subdued (e.g. the case of anger for the Utku, see Briggs (1970).

In some cases, culturally-specific emotional behaviors take the form of rituals, namely elaborate ceremonial acts. For example, one of the behaviors associated with anger among the Ilongot men is the ritual of communal headhunting, in which the angry individual looks for someone to behead jointly with his fellow tribesmen so as to vent his anger in ways that are not detrimental to the group (Rosaldo 1980). When the headhunting expedition is successfully

completed, the men return to the village and their return is celebrated by public songs. Among the Awlad'Ali, emotions such as sadness are expressed by the formulation of little poems read among intimates (Abu-Lughod 1986).

One of the most famous cases of culturally-specific emotional behaviors was first described by Newman (1964). Members of the Gururumba tribe, a community living in New Guinea, sometimes engage in the "wild pig" syndrome. The syndrome is a sort of ritualized anger, which the locals interpret as being caused by having been bitten by the spirit of a deceased person. The "wild pig" engages in a sequence of behaviors that are culturally specific in terms of their nature and duration. Such behaviors include looting objects of small value and shooting arrows, which rarely result in injury to anyone. These behaviors last for a few days, at the end of which the "wild pig" goes into the forest, where he spends a few more days before returning to the village in a calm state in which he claims not to remember anything about the episode.

The conscious experience of emotions is also likely to manifest cultural differences, as an effect of cultural differences in the other marks of emotionality. For example, in a culture with a display rule against crying (e.g. the Utku culture, Briggs 1970), sadness is likely to have a different feel, because it won't be accompanied by the physiological discharges that accompany crying. In a culture in which manifestations of happiness are disapproved of (e.g. the Ifaluk culture, Lutz 1987), happiness will also probably have a different feel. This is especially true because the bodily manifestations of emotions have been proven to affect their intensity. Studies have shown that sensory feed-back from facial and postural movements can intensify or reduce emotions, and in some cases even appear to cause them (see Izard, 1993).

The nature of the mental actions associated with emotions, finally, is also likely to manifest cultural variations. The acts of reasoning, memory and imagination generated in a certain emotional state will reflect the same cultural elements that shape the other marks of emotionality. For example, in a culture which disapproves of sadness (e.g. the Awlad'Ali, see Abu-Lughod, 1986), it is likely that sadness will be accompanied by thoughts of worthlessness, and mental plans for redressing one's wounded honor.

6.2.2. Do lexical emotion categories differ in different cultures?[9]

What kind of emotion taxonomies exist in the approximately 6,000 spoken languages other than English? The first important fact is that some languages do not collect subordinate emotion kinds such as fear, anger, disgust, shame, guilt and so on into a superordinate category such as our "emotion". For example, Tahitians (Levy, 1973, 271), Bimin Kuskusmin of Papua New Guinea (Poole, 1985), Gidjingali aborigines of Australia (Hiatt, 1978), Ifalukians of Micronesia (Lutz, 1980, 1983), Chewong of Malaysia (Howell, 1981), and Samoans (Gerber, 1975) appear to lack a term intertranslatable with "emotion". On the other hand, Samoans use the term *lagona*, which groups together emotions and sensations (Gerber, 1975). The Ifaluk speak of emotions as instances of *niferash*, i.e. "our insides".

Generally speaking, languages vary widely concerning their total number of emotion categories. Whereas estimates of emotion categories in the English language range from 500 (Averill 1975) to 2,000 (Wallace & Carson 1973), some languages are much less rich in this respect. For example, Lutz (1980) claimed to have found only 58 emotion categories in the Ifalukian language, and Howell (1981) detected only 7 emotion categories in Chewong. Levy (1983, 1984) argued that some cultures *hypercognize* certain emotions by generating large quantities of lexical categories for them. Conversely, other emotions are *hypocognized*, and very few lexical categories – sometimes none – are produced for them. For example, Levy (1983) argued that anger and sadness are respectively hypercognized and hypocognized by the Tahitians, who distinguish between 46 kinds of anger but lack the lexical category of sadness entirely. This does not mean that they do not experience sadness, but merely that that sadness is not categorized by a lexical item that designates it uniquely. Levy reports the Tahitian term *pe'a pe'a*, which does not distinguish between sadness, illness and fatigue, but is applied for example to situations of separation from loved ones.

It seems that every language L has at least some L-specific emotion category E, meaning that no language other than L has a word which designates the same combinations of evaluations, physiological reactions, expressions, instrumental and mental behaviors as E does. Since emotion categories are formed to serve purposes of communication on the background of a given social

[9] This section is a close summary of Russell (1991), where I found most of the references I cite.

context, this linguistic phenomenon is not surprising. Here are a few examples of language-specific emotion categories. Some can be understood as resulting from the precisification of antecedent circumstances of existing English emotion categories. The Czech language contemplates an emotion named *Litost*, which is "a state of torment caused by a sudden insight into one's own miserable self" (Kundera 1980, 121–122). The Japanese contemplate the category of *ijirashii*, which is a kind of joy caused by seeing that a commendable person overcomes an obstacle (Russell 1991). Ifalukians distinguish disgust caused by moral indignation, *song*, and disgust caused by detection of decaying matter, *niyabut* (Lutz, 1980, 183–184). They also distinguish fear of future events, *metagu*, from fear of present ones, *rus* (Lutz, 1980, 188).

Briggs (1970) reported that the Utku distinguish love towards vulnerable creatures, *naklik*, from love towards admirable people, *niviuq*. Morice (1978) detected fifteen kinds of fears in Pintupi, including *ngulu* (fear of another seeking revenge) and *wurrkulinu* (fear about land or relatives). The Japanese category of *amae*, first described by Doi (1973), designates a kind of joyful dependence towards someone, similar to what a child may experience towards a mother. In other cases, the L-specific emotion category is meant to introduce a new superordinate of existing English emotion categories.

For example, as described by Fajans (1983), Baining people of Papua New Guinea have the category of *awumbuk*, supposedly a combination of "sadness, lassitude, tiredness, and boredom caused by the departure of visitors, friends, or relatives" (Russell 1991). Shostak (1983) speaks of the !Kung having the category of *kua*, a combination of awe, respect, and fear caused by having one's life achievements publicly celebrated. Sometimes the L-specific category is meant to introduce a subordinate category of emotion which does not exist in English. For example, Lutz (1985) argues that the Ifaluk have the category of *nguch*, a kind of emotion roughly corresponding to our feeling sick and tired of something. *Fago*, discussed by Lutz (1980), is experienced when somebody dies or is in need, as well as when a gift is received or when one is in the presence of someone admirable, and it can be understood as a sort of combination of love, empathy, pity, sadness, and compassion.

In some cases, languages other than English lack English-specific emotion categories. For example, some African languages do not have a distinct category for anger and sadness (Leff 1973, 301). The distinction between shame and fear is not available to the Gidjingali aborigines of Australia (Hiatt, 1978), who use the category of *gurakadj* to cover both. Shame and

embarassement are blurred into one category by the Japanese (Lebra, 1983, 194), the Tahitians (Levy, 1973), the Ifalukians (Lutz, 1980, 209), the Indonesians (Keeler, 1983, 153), and the Newars of Nepal (Levy, 1983). The Ilongot do not have the categories of shame, timidity, embarrassment, awe, obedience, and respect, and use *betang* to cover them all (Rosaldo 1983, 141). The Javanese call *isin* the disjunction of shame, guilt, shyness, and embarrassment (Geertz, 1959, 233). Samoans use the category of *alofa* for love, sympathy, pity, and liking (Gerber, 1975, 3). The Utku do not distinguish kindness and gratitude lexically, calling both *hatuq* (Briggs 1970, 326).

In some cases, languages lack English emotion categories and do not even subsume them under more abstract categories. For example, Marsella (1981) argued that the category of *depression* is absent from many non-Western cultures. Eskimos and Yorubas lack the category of *anxiety* (Leff, 1973, 304). Johnson, Johnson, and Baksh (1986) did not encounter any category corresponding to our *worry* among the Machiguenga of Peru. Levy (1973, 342) argued that the Tahitians have "no word which signifies anything like a sense of guilt". Apparently, the category of guilt is also missing from the Sinhala language of Sri Lanka (Obeyesekere, 1981, 79), from the Ilongot language of the Philippines (Rosaldo 1983, 139–140), from the Pintupi language of aboriginal Australians (Morice 1978, 93), and from the Samoan language (Gerber, 1975). Gerber (1975, 3) stated that there is a "notorious absence of a term equivalent to guilt in many Asian and Pacific languages." The Quichua of Ecuador lack the category of remorse (Tousignant 1984), and apparently the Ifaluk lack the category of surprise. The Nyinba of Nepal do not have the category of love (Levine 1988), and refer to what we call love either in terms of compassion (for children) or in terms of desire (for sexual partners).

6.2.3. Does cultural variation support social constructionism?

The question to ask is: Are the cultural differences I described so far a good reason to say that the emotions are socially constructed? According to some social constructionists such as Armon-Jones (1986a, 33), the very fact that emotions are "elicited as a result of the agent having acquired a culturally appropriate construal of the situation" is sufficient ground to conclude that emotions are socially construed. Since appraisals are shaped by culturally specific material conditions, forms of self-understanding, shared concerns, values, and norms, we could say that

the emotions in a given culture "need not exist in their present form", and are not determined by the "nature of things".

This kind of argument can be reinforced by considering the influence of cultural factors on marks of emotionality other than appraisal. This is the line suggested by Averill, who argues that there are, besides social *rules of appraisal* which "pertain to the way a situation is perceived and evaluated", also social *rules of behavior*, which "refer to the way an emotion is organized and expressed", social *rules of prognosis*, which "concern the time course and progression of an emotional episodes", and social *rules of attribution*, which "pertain to the way an emotion is explained or legitimized" (107-108). Averill's convinction is that, even though "some of these component responses may be biologically based" – for example physiological and expressive responses – "the way the components are organized into coherent syndromes is determined primarily by social and not biological evolution" (1986, 100).

The evidence on cultural variations I have described is certainly fascinating. But is it enough *as such* to ground an interesting form of social constructionism? Affect program theorists are well aware of the fact that several features of the emotions are culturally specific. For example, Ekman and Friesen (1969, 73) have written that "[w]hile the facial muscles which move when a particular affect is aroused are the same across cultures, the evoking stimuli, the linked affects, the display rules and the behavioral consequences all can vary from one culture to another".

Even though Ekman discussed mostly culturally-specific *display rules* for expressions, the same rationale that led him to posit such rules should motivate positing culturally-specific appraisal rules, behavioral rules, assessment rules and so on, as suggested by Averill. But the presence of such rules is no threat to the affect program position. The affect program thesis *is not* that all marks of emotionality in all emotions are universal across cultures, but rather that there exist *some* marks of *some* emotions that are universal, and thereby support the hypothesis of evolutionary origin. Holding that some parts of emotions are universal is perfectly compatible with holding that some others are culturally-specific.

The ethnographic evidence does not offer us any reason to think that there are no universal elicitors, expressions and behavioral tendencies for emotions such as anger, fear, surprise, joy, sadness, disgust. For example, a car accident will surely elicit the same fear response in all cultures, namely the same automatic elicitation of an affect program with several of the features described by Ekman. On the other hand, the evidence I have described points us to a limitation of

affect program theory I have already remarked upon, namely that it fails to comprise within its purview a large bulk of affective phenomena that in ordinary language we call emotions. Several emotion episodes do not have distinctive universal signals, their appraisal is not automatic nor likely to be mediated by subcortical pathways, their antecendents are not universals, they do not occur in other primates, their onset is slow, they last for longer than a few seconds, and so on. In other words, affect program theory is incomplete as a theory of emotions as ordinarily understood. But this is not something affect program theorists are likely to deny. Their rationale for focusing on basic emotions, as I have argued, is that they assume them to share deep similarities stemming from their evolutionary origin. Moreover, the basic emotions approach, with its emphasis on observable characteristics, makes the scientific study of the emotions possible.

In conclusion, the version of social constructionism embodied by the *emotions as culturally-specific syndromes thesis* seems to me to violate Hacking's (1999) pre-condition (0). Since emotion theorists of all stripes accept the presence of cultural variation, there is no room for consciousness-raising with respect to this issue. The task for social constructionists is a different one, namely to explain in what respects the social dimension of emotions is crucial to understanding them. For example, what is the importance of the social dimension of emotional phenomena to understand their evolution? Are there aspects of the emotions other than those stemming from their evolutionary origin that make them proper objects of scientific investigation? Can the study of the social dimension of emotions move from mere speculation to scientific test? The core of an interesting social constructionism, and the springboard for answering these sorts of questions, lies in my view in the *emotions as roles/transactions* thesis, to which I now turn.

6.3. EMOTIONS AS SOCIAL ROLES AND INTERPERSONAL MOVES

This strand of social constructionism questions the view of the emotions as *passions*, pointing to the existence of ends served by the emotions in light of which they should rather be understood as actions or roles or moves. The assumption that the emotions are passive is as old

as the study of emotion itself, and it is embodied by the very name under which the emotions have been known for most of their intellectual history, starting with the Aristotelian theory of *pathe* (see subsection 2.1.1) In this sense, this version of social constructionism unquestionably fulfills Hacking's (1999) precondition (0), as the large majority of emotion theorists have historically taken and currently take the passivity of the emotions for granted. Ekman (1999b, 54), for example, argues that "[b]ecause emotions can occur with a very rapid onset, through automatic appraisal, with little awareness, and with involuntary changes in expression and physiology, we often experience emotions as happening to us. Emotions are unbidden, not chosen by us".

This strand of social constructionism puts pressure on the very idea that emotions are not chosen by emoters. Notice that this version of social constructionism is conceptually distinct from the previous one, even though it is generally conjoined with it. Even if the emotions did not change across cultures, they may still have to be understood in each culture as roles/transactions rather than passions insofar as they fulfill ends (the ends may be the same in every culture).

We can distinguish two versions of the *emotions as roles/transactions* thesis. According to the first, proposed most prominently by Averill (1980), the emotions fulfill ends at the societal level, solving conflicts between norms that regulate behavior in society. Call this the *emotions as social roles* thesis. According to the second version, proposed most prominently by Parkinson (1995) and Griffiths (2003, 2004a), the emotions fulfill personal ends emerging in the course of a social transaction. Call this the *emotions as social transactions* thesis. This second approach comes with a strong emphasis on the strategic nature of the emotions, so I will deal with it only after having discussed the emergence of the idea of a strategic dimension in emotional phenomena.

6.3.1. Sartre

The importance of Jean-Paul Sartre for the social constructionist approach to the study of emotions is that he is the first to offer a theory of emotions as strategic moves in the negotiation of a social transaction. Jointly with evidence on cultural differences, this is the main idea that motivates social constructionism about the emotions. Sartre's version of ante litteram social constructionism is very radical, and often expressed in a characteristically obscure language. Nevertheless, once we clarify what Sartre meant and purge his theory of some of its flamboyant

non-sequiturs, a number of useful insights, later independently developed by others, can be gleaned.

Sartre (1948) offered his theory as a reaction to the perceived shortcomings of James-inspired theories of the emotions. His main complaint was that such theories focused on "the processes of the emotion itself", rather than on the "general and essential structures of human reality" that make emotion possible (9). This approach is problematic because emotion as studied by psychologists "will never be anything but a fact among others", and will not permit "grasping by means of it the essential reality of man" (9). According to Sartre, what we need to do is to "study emotion as pure transcendental phenomenon, trying to "elucidate the transcendental essence of emotion as an organized type of consciousness" (12). Sartre states that "every consciousness exists to the exact extent to which it is conscious of existing" (11). What follows from this is the "absolute proximity of the investigator and the thing investigated".

Phenomenology, states Sartre, "is the study of phenomena, not facts", where a phenomenon is "that which manifests itself", or "that whose reality is precisely appearance" (14). For consciousness, "to exist is to appear" (14). Sartre cites Heidegger and Husserl as early proponents of the right approach to the study of emotions. Heidegger thinks that "emotion is the human reality which…directs itself towards the world", and Husserl adds that "a phenomenological description of emotion will bring to light the essential structure of consciousness, since an emotion is precisely a consciousness" (15). The fundamental ground of difference between thinking of emotions as *facts* and thinking of them as *phenomena* is for Sartre that, whereas psychologists conceive of emotion as a *fact from which signification has been removed*, for the phenomenologist "every human fact is, in essence, significative" (16).

Consequently, in order to study emotion as a "true phenomenon of consciousness", we have to understand emotion as something "significative", where to "signify is to indicate another thing" (16). What Sartre invites us to consider is that "emotion signifies, in its own way, the whole of consciousness and…human reality" (17), since "emotion is an organized form of human existence" (18). The methodological consequence of these remarks is that, in order to study the emotions, we must interrogate "phenomena, that is, to put it exactly, psychic events, insofar as they are significations and not insofar as they are pure facts" (19). It is at this very juncture that the contrast between Sartre's theory and the feeling theory of emotions emerges most compellingly.

Sartre is convinced that thinking of emotions *as significant phenomena of consciousness*, as recommended by phenomenology, is incompatible with thinking of emotions *as physiological facts*, as recommended by feeling theories of Jamesian ancestry. As he puts it, "physiological facts…taken by themselves and in isolation…signify almost nothing" (17), and therefore "it is impossible to consider emotion as a psychological disorder" (17). As Sartre puts it, in the case of emotions we have "signification by nature of a functional order" (41). The functional order of the emotions consists of their being "an organized system of means aiming at an end" (32). What is special about the emotions is that they aim at their ends in a *masked* or *covert* fashion. Writes Sartre:

> [Emotion] is called upon to mask, substitute for, and reject behavior that one cannot or does not want to maintain. By the same token, the explanation of the diversity of emotions becomes easy; they represent a particular subterfuge, a special trick, each one of them being a different means of eluding a difficulty (32)

Sartre's central idea is that we emote when the opportunity to *pursue our ends* in non-emotional ways turns out to be unavailable or unappealing. Under this view, we emote by *substituting* a behavior that is *openly* instrumental with one that is *covertly* instrumental. Sartre labels the non-emotional instrumental behavior as the *superior* one, and the emotional instrumental behavior as the *inferior* one. We can better understand what Sartre had in mind by focusing on a few of his examples. Sartre describes the case of "a young girl whose father has just told her that he has pains in his arms and that he is a little afraid of paralysis". At this point, the girl undergoes a sequence of violent emotions, which finally lead her to see a doctor. In the course of treatment, "she confesses that the idea of taking care of her father and leading the austere life of a sick nurse had suddenly seemed unbearable" (26). In such case, the *superior behavior* would have been to *openly* reject the role of nurse, and accept the social disapproval that would ensue.

By having an emotional break-down, the girl *covertly* rejected that nursing role, resorting to an *inferior behavior* to achieve that objective. Sartre's insight is that we should not merely say that the inferior behavior *substitutes* the superior one, but rather that it occurs *in order to substitute* the superior one. This is what gives emotions their *covert finality*. Sartre offers a variety of suggestive examples of this sort of finality in the emotional domain. He reports the case of a woman who had begun an embarrassing confession to Pierre Janet, one of the first doctors to recognize and treat psychosomatic illnesses in 19th century France. All of a sudden,

the girl interrupted her confession, and started to sob copiously. In such case, emoting was a means to cope with "a narrow and threatening world which expected her to perform a precise act", and it resulted in achieving the end of dissolving the "intolerable tension" generated by the confession, thereby "transforming Janet… from a judge to a comforter", and "canceling the precise necessity to give such and such information" (38).

Sartre also reports his own experience of getting angry when "I could not longer reply to someone with whom I had been bantering" (38). In such case, the *superior* behavior would have been to reply with another witticism, but lacking such option for the inability to come up with a witticism, Sartre resorted to getting angry as a means "to conquer my opponent". Sartre takes the covert finality of emotions to be the *essence of emotion*, an essence which becomes apparent to anyone willing to make an "objective examination of emotional behavior" (41).

We should not construe Sartre's notion of a subterfuge, I suggest, in terms of a deliberate pretense. Sartre remarks that there is no "reasoned calculation" (36) in the covert finality of emotion. As he puts it, "emotion is a certain way of apprehending the world". Superior behaviors, what we would call standard actions, are for Sartre the result of an "apprehension of the means as the only possible way to reach the end". When "the paths traced out become too difficult, or when we see no path", standard actions must be substituted with something else. It is at this point that "we try to change the world, that is, to live as if the connection between things and their potentialities were not ruled by deterministic processes, but by magic" (59).

An emotion can therefore be understood as "a transformation of the world" (59) according to the laws of magic. "There is emotion", Sartre tells us, "when the world of instruments abruptly vanishes and the magical world appears in its place" (90). Importantly, the emoter's "attempt [to transform the world] is not conscious of being such, for it would then be the object of a reflection" (59). I do not interpret Sartre as saying that emotions are not access-conscious in Block's (1995) sense (see 4.1.1). What I take him to be saying, rather, is that people do not accompany their access consciousness with respect to being, say, angry or sad with the intention of achieving the ends that their being angry or sad *does* achieve.

Sartre gives several further examples of the ways in which emotions serve masked ends. For example, "sadness aims at avoiding the obligation to seek new ways", and it does so by *magically* turning the universe into a gloomy place in which one must no longer cope with the difficult situation at hand (65). Joy is "a magical behavior which tends by incantation to realize

the possession of the desired object as instantaneous totality" (69). Fear "is a consciousness which, through magical behavior, aims at denying an object of the external world" (64). For example, one may encounter a dangerous animal, experience fear and faint, thereby finding refuge in the denial of danger, rather than in the adoption of a behavior of escape that is no longer available. The language of *magic* and *incantations*, as the language of *subterfuges* and *tricks* before, is an example of Sartre's penchant for flamboyance over clarity.

What Sartre aims to say is that the special consciousness needed for the instantiation of the masked finality of the emotions has an important physiological underpinning. Sartre thinks that the emotions generate their magical world "by using the body as a means of incantation" (70). An "emotion [is] an abrupt drop of consciousness into the magical" (90), caused by the special role played by the body in emotional episodes. Sartre compares the *magical consciousness* into which we drop by emoting to "the consciousness of sleep, dream, and hysteria" (77). He writes that "[i]t is necessary to speak of a world of emotion as one speaks of a world of dreams or of worlds of madness" (80). The analogy, as I understand it, is due to the fact that in such states one is inclined to take the world to be as it appears, as we are inclined to do when the world appears to us as scary, infuriating, or sad.

Sartre does not say much about magical consciousness, but he is clear that the magic of emotions comes from their physiological underpinnings. In emotion, he tells us, consciousness "is caught in its own trap [as] it lives the new aspect of the world by believing in it" (78). The *trap* is the one predisposed by the emoter's own physiology. "In order for us truly to grasp the horrible", says Sartre, "we must be spell-bound, flooded by our own emotion" (74), a state we reach by virtue of "purely physiological phenomena [which] represent the seriousness of the emotion" (74). He remarks that "*in order to believe* in magical behavior it is necessary to be highly disturbed" (75). This clarifies that Sartre's primary complaint with respect to James' theory of the emotions is not that it reduces consciousness to its physiological underpinnings, but rather that it considers such physiological underpinnings *in isolation*.

Sartre wants us to understand the *physiological* in emotional phenomena *in relation* to their *masked finality*. The emotion is "the demeanor of a body which is in a certain [physiological] state; the state alone would not provoke the demeanor; the demeanor without the state is comedy…without [the] disturbance, the behavior would be pure signification, an affective scheme" (75). The body *resonates* in emotional phenomena in such a way that the

emoter is *truly struck* by the horrible, the joyful or the disgusting, and interacts with the world in light of the acceptance of its emotionally apprehended qualities. Such acceptance need not be reflectively underwritten, a point suggests when he compares *emoting* to *hallucinating* or *dreaming*.

At the end of his book, Sartre (1948) himself seemed to realize that there was at least one case in which his account of masked finality did not work well. He pointed out that "this theory of emotion does not explain certain abrupt reactions of horror and admiration which appear suddenly", as when we unexpectedly see a "grinning face…flattened against the window pane" (82). In such cases, Sartre states that "it seems that the emotion has no finality at all" (83). But he thinks that the general spirit of his theory can be maintained, if we understand the case of sudden horror as one in which "the world itself…reveals itself to consciousness as magical" (83). The idea here is that in the case of horror directed towards the grinning face, we do not *project* the horrible into the world as if by magic, but rather the horrible *manifests itself* as an "existential structure of the world which is magical".

The distinction between a projected and an independently existing magic is not well worked out, and Sartre does not do much to shed light on it. His suggestion is that "there are two forms of emotion, according to whether it is us who constitute the magic of the world to replace a deterministic activity which cannot be realized, or whether it is the world itself which abruptly reveals itself as being magical" (85). He concludes by saying that "there are often mixtures of the two types and most emotions are not pure" (86). With this final caveat, Sartre takes himself to have provided a fully general theory of the emotions that achieves what it aimed for, namely restoring a central role for the *psychic*, and opposing an understanding of the emotions as *mere* physiological disturbances.

The *psychic* had been restored because emotion had been shown to be a phenomenon of *magical consciousness*, and the *physiological* component had been accounted for as an essential part of the "total modification of "being-in-the-world" according to the very particular laws of magic" (93). Sartre's assimilation of reflex emotions to the pattern of finality typically instantiated by crying in order to be comforted is unprincipled, and ought to be rejected. More generally, we should reject the presupposition that a *single account* of the finality of emotions will explain all forms of finality instantiated by emotional processes. Sartre is refreshingly sensitive to the ends our emotions allow us to achieve by virtue of what they *communicate* in the

course of an ongoing social transaction. For example, being sad communicates the need for comfort, and Sartre reveals remarkable insight when he suggests that understanding sadness demands, among other things, understanding how the evaluation of the payoff to be gained by getting sad causally impacts on the appraisal by virtue of which we get sad. But he is wrong if he assumes that every case of finality is somehow derivative from the general model he has provided.

The case of reflex emotions, to elaborate on Sartre's own example, seems particularly problematic for an account that assumes the emotions to occur when, after having considered the situation at hand, we realize that "the paths traced out become too difficult, or…we see no path", and thereby substitute a standard action with an inferior emotional behavior. Reflex emotions are elicited *before* the presence or absence of *paths* can even be considered, and they do not seem to involve any sort of strategic evaluation. As I see it, the value of Sartre's account of finality does not lie in its generality, but in the fact that it points our attention to the *strategic* dimension of communication in emotional phenomena.

Sartre's own way to reconcile strategic and non-deliberate dimensions of emotionality was not entirely satisfactory, but he must be commended for having questioned with unprecedented force the commonsensical view that emoting is a way of being acted upon, pointing out various ways in which emoting is in fact more like acting. This insight, as well shall see, has been developed into the important idea that emotions have a Machiavellian dimension which is crucial to understand their origin and current function (Griffiths 2003, 2004a).

6.3.2. Averill

Let us now consider Averill's *emotions as social roles* thesis. Its inspiring thought is that "emotions are responses that have been institutionalized by society as a means of resolving conflicts which exist within the social system" (1980, 37). The evidence I discussed in the section on cultural differences offered us examples of the *institutionalization* of emotional responses, namely of the fact that each of the marks of emotionality is affected by culturally-specific rules of appraisal, of display, of behavior, and so on. The key addition here is that such institutionalization is *aimed at solving societal conflicts*, under the presupposition that "[f]or a response to become institutionalized, it must serve some social function" (1980, 47). In this

sense, the emotions are *social roles*, where a social role is "socially prescribed set of responses to be followed by a person in a given situation" (Averill 1980, 308).

But clearly not all social roles generated to solve social conflicts count as emotions. For example, the social role of a policeman, which comprises a socially prescribed set of responses aimed at resolving societal conflicts, is something other than an emotion. The final piece of Averill's account is that the social roles that are emotions are on the one hand *temporary* and on the other hand *construed as passions*. As Averill (1980) puts it, "[a]n emotion is a transitory social role (a socially constituted syndrome) that includes an individual's appraisal of the situation, and is interpreted as a passion rather than as an action" (139). The *rules of attribution* by virtue of which emotions are interpreted as passions are key to distinguishing social roles such as policeman from social roles such as anger.

Averill argues that the assumption of passivity results from a limited understanding of the social role played by emoting. As he puts it, "the experience of passivity during standard emotional reactions presume[s] a limited self-awareness, a restricted insight into the sources of one's actions" (1980, 65). Under this view, the emoter does not feign anger, love, grief, sadness and so on, but rather goes through such emotions without fully understanding that he or she is engaging in them in order to resolve existing conflicts. Averill believes the "wild pig" syndrome I described in the previous section (Newman 1964) offers a general blueprint for what the emotions (as social constructions) are. Notice that this represents a diametrically opposed choice of paradigm case with respect to affect program theorists, who take the brief episode of fear caused by loss of support to be a general model for what the emotions (as affect programs) are.

As I reported, people in the Gururumba society sometimes become "wild pigs", and engage in a ritualized series of aggressive acts which last a few days, after which they hide in the forest to come back a few days later allegedly without recollection of what they have done. Averill suggests that the social function of the syndrome is to give people "an acceptable means of communicating the difficulty" (1980, 47) encountered in dealing with social responsibilities. The syndrome only occurs in people between the ages of 25 and 35, a critical age in the life of a Gururumba male. This is in fact the time in which males get married and begin sharing a significant portion of communal responsibilities. When the wild pig man returns from the forest in a state of calm, nobody reproaches him for his rampage and the members of the tribe

"reevaluate the individual's ability to meet his obligations, and expectations are adjusted accordingly" (1980, 45).

Importantly, neither the "wild pig" nor the community around him have a full understanding of the wider social context in which the emotional episode is embedded. As I pointed out before, within the Gururumba society it is assumed that what causes the syndrome is "being bitten by the ghost of a person who has recently died" (1980, 44). This is a vivid example of what it means to interpret an emotion as a passion, namely as something that occurs unbidden, without insight into the larger social dynamics that bring it about.

But can this account be generalized to all emotions? Averill (1980) is convinced that "most standard emotional reactions…are behaviors which frequently are condemned by society and yet which are maintained because they serve useful (but disguised functions) within the sociocultural system" (66). We have a further thesis here, namely that emotions are not only means of resolving a societal conflict, but also that the societal conflict concerns the very kinds of behaviors comprised in the syndrome. Under this view, emotions emerge when there are "norms which simultaneously encourage and discourage a particular kind of behavior", which becomes an emotional behavior to be justified in the name of passivity.

For example, anger is the kind of behavior discouraged by norms against violence and encouraged by norms in favor of protecting one's own rights from infringers. By being "overcome" by anger, individuals manage to protect their rights by inflicting violence, and are justified in so doing so because anger allegedly overcame them (Averill 1980, 66). An obvious problem with this account is that a large number of emotions consist of behaviors that *either* are neither encouraged nor discouraged *or* are either encouraged or discouraged, but not both.

Love behaviors seem to be a good example of behaviors that are generally encouraged but nor discouraged, whereas there do not seem to be specific social norms about, say, surprise behaviors. Moreover, in the case of many emotions there are no responses that are prescribed, namely such that one has an obligation to perform them. For example, there are very many loving behaviors one may choose when in love, but the presence of a unique ritualized sequence as in the wild pig case appears absent. We may substitute prescribed with "prescribed or permitted", so as to enlarge the domain of emotions understood as institutionalized responses in a broader sense.

The real problem, however, concerns the assumption of *conflict resolution*. Averill is forced to make up sui generis conflicts to bestow generality upon his account. He proposes for example that romantic love is discouraged by norms which require the anonymity of individual subjects, and encouraged by norms demanding the promotion of self-worth. This is clearly a stretch, which shows that there is something wrong with the idea that emotions always come about to solve societal conflicts, more precisely conflicts regarding emotional behaviors themselves.

I think Averill has missed the main insight offered by rejecting the idea of emotions as unbidden occurrences. His account presupposes that what one does by emoting is essentially performing an act that is societally prohibited in a way that becomes justifiable because of the assumption of passivity. This view suggests that the point of emoting is getting away with a norm violation. On the contrary, it seems to me that if there is a point to emoting by virtue of which we can assimilate emoting to an activity, it is that of bringing about certain *effect*s in the context of a social transaction. This is the thesis of *emotions as social transactions*. Bringing about such effects may not involve behaviors discouraged by social norms, and even when it does, the point of emoting is bringing about the effects, not getting away with the norm violation.

Consider anger once again. According to Averill, anger serves the purpose of making aggressive behavior justified. The interpretation of passivity here results from the emoter lacking insight into the fact that he is emoting so as to avoid the social disapproval that would be associated with his behavior had he not been overcome by it. The problem is that people don't get angry in order to be justified in behaving aggressively, but in order to obtain something *by* behaving aggressively. As first suggested by Sartre, sometimes people get angry in order to "conquer an opponent", namely in order to bring about a submissive response. The interpretation of passivity here results from the emoter lacking insight into the fact that he is emoting so as to bring about a certain response in the opponent.

What is masked is *not* the *resolution of a social conflict* between norms, but the *pursuit of a personal end*. The same disregard for the personal ends emoting may fulfill in the context of a personal transaction appears in all accounts of emotions as social roles mentioned by Averill. For example, falling in love while already married is much more likely to serve the purpose of changing sexual partner while reducing the blame connected to violating a bond of trust ("I could not help it, I'm a sorry"), than the purpose of getting away with violating the (alleged) social norm discouraging anonymity.

Similarly, the wild pig in the Gururumba society is not likely to engage in the syndrome in order to be justified in shooting arrows and looting other people's properties, but in order to have his social obligations reassessed. This leads me to conclude that the most promising version of the *emotions as roles/transactions thesis* takes the emotions to be social transactions in the sense of *interpersonal moves*.

But what is the evidence that emotions are social transactions? Aren't emotions the arena of the automatically elicited and unbidden? Don't emotions occur so quickly that there simply is no time for strategic considerations? Aren't facial expressions involuntary and impossible to reproduce at will?

6.3.3. Hinde and Fridlund

The picture of emotions emerging from the Darwin-Tomkins-Ekman tradition has the following distinguish properties. An automatic appraisal is assumed to occur in correspondence to universal antecendents. The appraisal is followed by a cascade of responses, including physiological, expressive and behavioral ones, which follow quickly and mandatorily from the appraisal. A specific expression is universally associated with each basic emotion, and consequently it carries veridical and highly reliable information about what emotion is unfolding. Under this view, expressions are *recipient-independent*, and *informative* of the occurrence of a specific affect program. The expressions produced by the affect program, on the other hand, can be partially inhibited or modified - display rules are at work – and they can be faked. Ekman (1972) argued that in the first case, the universal expression is likely to transpire, whereas in the second case it is not likely to be accurately reproduced.

A rather different picture of emotional expressions has emerged in the last fifteen years, commonly associated with the names of Fridlund (1997) and Russell (1997), but importantly anticipated by Hinde (1985a, 1985b). As I understand it, their work suggest that at least some emotional expressions are *recipient-dependent* and *strategic*, in the sense that they are specifically directed towards a recipient, and they do not necessarily inform him of the occurrence of an affect program. Rather, they aim to generate an effect in the recipient that is advantageous to the sender. Whereas the Darwin-Tomkins-Ekman tradition thinks of expressions primarily as external read-outs of an occurring emotion (call this the *traditional view*), the

Hinde-Fridlund-Russell tradition thinks of expressions primarily as moves in the course of an ongoing negotiation (call this the *behavioral ecology view*).

The ethologist Robert Hinde (1985a, 1985b) was the first to make an articulate case for the negotiating role of expressions. He began by noticing that several kinds of expressions, in both humans and animals, are issued only when a recipient is there to be influenced by them, and that the responses of the recipient to the expression determine what behaviors follow it. For example, Hinde pointed out that in several cases birds flee after having issued a threat expression (Stokes 1962). His interpretation was that threat displays "were given when the bird was uncertain what to do" and that "which of the several possible responses it showed next depended on the behavior of the rival" (Hinde 1985a, 109). Under this view, threat expressions should be understood as "signals in a process of negotiation between individuals".

These ideas were applied by Hinde (1985a) to cast doubt on the assumption that "emotional behavior is the outward expression of an emotional state, and that there is a one-to-one correspondence between them", an assumption Hinde associated with Darwin. By emotional behavior, Hinde understood mostly facial and postural expressions. Hinde thought that the assumption of a one-to-one correspondence does not make evolutionary sense, as it may be adaptive for an organism to mislead the recipient of the expression about the nature of the internal state to which it is associated (Dawkins and Krebs 1978). In many cases, natural selection will favor sending non-veridical or ambiguous messages, and take a course of action contigently upon the response received.

Hinde, however, acknowledged that signals do not always serve negotiating purposes. For example, in the case of the "begging calls of a young bird" the signal is indeed the read-out of an internal state. Hinde's (1985b) conclusion was that we should expect expressions to lie on a *continuum* between *expressing* and *negotiating*. He wrote:

> Such considerations suggest the view that emotional behavior may lie along a continuum from behaviours that is more or less purely expressive to behavious concerned primarily with a process of negotiation between individuals…In animals, bird songs lies nearer the expressive end, threat postures nearer the negotiation end. In man, spontaneous and solitary laughter are primarily expressive, the ingratiating smile primarily negotiating. However most emotional expressions involve both (Hinde 1985b, 989)

I consider this to be an important insight. What I take Hinde to be suggesting is that most emotional signals have a non-arbitrary relation to the states to which they are associated but at the same time are aimed at making a move in a negotiation whose outcome is open-ended and crucially dependent upon the recipient's responses. In chapter 10, this idea will be put at the center of the theory of emotions I propose.

Ever since Hinde wrote on this issue, much research has been offered in support of the view that emotional expressions have a strategic dimension. The debate has become quite radicalized, and the idea of a continuum has gone lost in favor of views that take expressions to be either pure read-outs or pure negotiating moves.

Fridlund (1989, 1997) is the best known contemporary researcher on the strategic dimensions of emotional expression. The basic tenet of his view is that facial displays are not "'expressions' of discrete, internal emotional states, or the outputs of modular affect programs" and that "displays have their impact upon others' behavior because vigilance for and comprehension of signals coevolved with the signals themselves" (Fridlund and Duchaine 1996, 278). This view is accompanied by skepticism on the universality of facial expressions, articulated most prominently by Russell (1994) (see Ekman 1994 for a response). Some of the points of contention have been surveyed in my review of Ekman's experiments in subsection 5.2.4, so I won't discuss them again here.

Fridlund and Duchaine's (1996) view is that "there may be no prototypes [sic] faces for each category. Rather, displays exert their influence in the particular context of their issuance" (278). Under this view, there is no expression universally associated with each of the basic emotions. Anger expressions, for example, will comprise all the expressions which can be used in different circumstances to convey something like the message "I am likely to attack you". This is not simply to say that the universal expression of a basic emotion can be inhibited or transformed, as Ekman would have it with his proposal of display rules. Rather, the idea is that no universal expression is automatically associated with basic emotions at all.

For example, according to the *traditional view*, Duchenne smiles are always generated as part of the affect program of happiness. Fernández-Dols & Ruiz-Belda (1997b) demonstrated that Olympic gold medal winners, allegedly a paragon of happiness, generate the Duchenne smile only when interacting with other people. We can interpret this fact as indicating that smiles are not broadcast in a recipient-independent way as part of an automatic cascade of responses,

but are issued instead in order to negotiate a social transaction. The same is true of the smiles of soccer fans on the occasion of a goal of their favorite team. As shown by Fernández-Dols & Ruiz-Belda (1997b), Duchenne smiles are only issued when fans face one another.

Manifestations of the influence of recipients on expressions - the so-called *audience effects* - have also been shown with respect to the expressions generated by tasting and smelling and by pain (see Russell et al. 2003 for a recent review of the literature on audience effects). The behavioral ecology view, it seems, faces a difficulty in light on the fact that expressions are sometimes issued when the emoter is alone. Many take this to be a clear sign that the behavioral ecology view cannot be a general theory of expressions, as it covers at best a subset of the domain of expressions. For example, Ekman (1990, 263) wrote that "[f]acial expressions do occur when people are alone…and contradict the theoretical proposals of those who view expressions solely as social signals". Hinde himself considered solitary laughter to be an example of the purely expressive nature of some emotional expressions.

Fridlund (1990) has argued instead that we should not infer that audience effects are absent just because an audience is physically absent. Audience effects may be generated by *imagining* an audience. This is an old idea in the literature on emotional expressions, initially formulated by Piderit and Gratiolet. Piderit wrote in 1858 that "the muscular movements of expression are in part related to imaginary objects, and in part to imaginary sensory impressions" (Piderit as quoted by Fridlund and Duchaine 1996, 265). The problem is that just positing the presence of imaginary audiences when expressions appear not to be social signals risks turning the hypothesis into an unfalsifiable one.

Fridlund (1990) realized this risk, and tried to test the imaginary audience hypothesis experimentally. He first proved that experimental subjects smile much more often when viewing a humorous videotape in the company of a friend than when alone. To test the hypothesis that audience effects can be implicit, he measured the amount of smiling in experimental subjects who watched the videotape alone but under the experimentally induced assumption that their friend was watching the same videotape at the same time in another room. If audience effects depended exclusively on the physical presence of an audience, there should be no difference in smiling between this situation and the alone viewing condition. But Fridlund (1990) documented that people smile almost as much when their friend is actually present as they do when they just

imagine him to be engaged in the same activity. The experiment does seem to show that we cannot reliably infer the absence of audience effects from the mere absence of an audience.

Although the literature on the social dimensions of expressions has been dismissed by some (e.g. Ekman 1994), I am convinced it offers an important insight on the nature of the emotions. But I think it should be integrated with, rather than opposed to, the literature on universal emotional signals. What the data offered by Fridlund and others make a case for is that in many cases emotional expressions work as negotiating signals. The evidence, however, does not prove that emotions never have automatic expressions associated to them, nor that such expressions are not universal. The facial expression of someone about to have a car accident, or of someone suddenly losing support, is going to be automatic and universal, and it does appear to be purely expressive.

Incidentally, Fridlund's (1990) own experiment shows that the idea that emotions are never purely expressive is problematic, as subjects in the alone condition *do* smile in the absence of an imagined audience. If smiling in the alone condition while imagining an audience suggests that the physical absence of an audience is no proof that audience effects are absent, smiling in the alone condition while not imagining an audience suggests that audience effects do not offer a complete explanation of facial expressions.

Some of the examples of emotions used by Fridlund and his associates, moreover, do not appear to qualify as affect programs. For example, the happiness of an Olympic champion, which behavioral ecologists assume not to be automatically associated with Duchenne smiles, is an extended episode which certainly lasts for more than a few seconds, and does not appear to be governed by an automatic appraisal mechanism. The data on the negotiating aspects of the happiness of Olympic champions should probably be constructed as evidence that many ordinary episodes of happiness are not affect programs, rather than as evidence that affect programs lack universal expressions.

There is another, more important reason why it is better to avoid the dichotomy between purely expressing and purely negotiating when it comes to signals. If there were no reliable connection between displays and internal states, understood as causes of ensuing behaviors, displays could hardly function as signals at all. For example, if threat displays were associated with equal probability with aggressive and peaceful behaviors, their negotiating potential would vanish. To say that signals are at the same time expressive and negotiating, as suggested by

Hinde, hints at the necessity for signals to be connected with internal states by a reliable, if not one-to-one, correspondence.

I propose we concede to Ekman that a handful of emotional episodes are automatically associated with recipient-independent and universal expressions, but point out that this is an incomplete theory of facial expressions. Many emotional expressions of many emotions have negotiating dimensions we cannot neglect if we want to understand their evolutionary origin and current function.

6.3.4. Parkinson and Griffiths

If emotional expressions are at least in part negotiating *signals*, couldn't emotions as a whole be at least in part negotiating *moves* in an ongoing *social transaction*? This is precisely what the *emotions as social transactions* thesis assumes to be the case. I will call *transactionalist* any emotion theorist who endorses this thesis. Brian Parkinson (1995, 2004, 2005) is the most prominent transactionalist in psychology, and Paul Griffiths (2003, 2004a) has began to show what philosophical dividends can be paid by this emerging research program. One difference between them is that, whereas Parkinson tends to contrast affect program theory with the thesis that emotions are interpersonal moves, Griffiths takes this thesis to complement the existing literature on basic emotions.

I pointed out in the last subsection that signals in many cases are not likely to be recipient-independent effects of internal states, but rather signals meant to influence a signal recipient in an advantageous way. What transactionalists want to make a case for is that emotions themselves in many cases are not interactant-independent effects of appraisals, but rather moves meant to influence an emotion interactant in an advantageous way. The hypothesis is that emotions may be *produced*, rather than merely *expressed*, in a strategic way.

Parkinson's (1995) starting point is the view that "[a]lthough certain aspects of emotional response derive fairly directly from our biological heritage,…I do not believe that it is possible to give a full account of all human emotional states by exclusive reference to evolutionary explanations" (163). Parkinson's view is that emotions "are social constructions, and they can be fully understood only on a social level of analysis" (1995, 309). Griffiths (2003) considers instead the social level of analysis to complement the evolutionary one. As he puts it, "a socially-oriented ('Machiavellian') perspective on the basic emotions can be incorporated into a theory of

extended emotion episodes…in such a way as to provide biological underpinnings for ideas that have traditionally been associated with social constructionist…accounts of emotion" (2003, 49).

I side with Griffiths in thinking that an evolutionary account is entirely compatible with one that emphasizes the strategic and social dimension of emotional phenomena. Hinde (1985a, 1985b) and Fridlund (1990, 1997) offer an example of this integration, as they diagnosed the emergence of strategic aspects as resulting from evolutionary pressures. Also, I argued that the very distinction between the social and the biological is problematic under an interactionist approach to the emotional phenotype such as the one I endorse. What Parkinson and Griffiths have in common is that they point us to aspects of the emotions that have been long neglected in emotion theory.

The key difference between the *emotions as social roles* thesis endorsed by Averill and the *emotions as social transactions* thesis endorsed by Parkinson and Griffiths is that according to the latter thesis what makes emotions goal-directed is not that they aim to solve societal conflicts but rather that they aim to promote the emoter's interests. The promotion of social ends, if there is one, is a secondary phenomenon under the transactionalist view. The primary phenomenon is that emoters emote in order to influence those they interact with in an advantageous way. This contrast is remarked upon by Parkinson (1995):

> [M]any of the occasions for emotion arise from local negotiations in the course of everyday personal interaction and do not directly reflect societally prescribed norms. The conflicts, disagreements, and commitments that lead to emotion may be based on mutually established rights and obligations in relationships which have only a remote connection with culturally imposed rules and roles (162).

What is added to this distinction is a *dynamic dimension* neglected by the emotions as social roles thesis. Averill characterized emotions as "responses that have been institutionalized by society", describing the production of such responses as akin to the performance of a script (although without insight, on the emoter's part, that it is a script). The model is that of a behavioral *ritual*, as the one manifested by the wild pig in the Gururumba society (see subsections 6.2.1 and 6.3.2) or by someone who expresses grief at the funeral by going through a fixed sequence of responses (e.g. crying at a particular time, fainting at another, etc.).

This view misses the point that emotional episodes often result from a dynamic give-and-take, in which responses are not pre-ordained at the beginning of the sequence, but rather shaped by the interactant's responses. This is where the metaphor of *negotiation* comes to full fruition, as the emotional episode is not exhausted by the interactant's reception of a one-shot message, but is rather dynamically shaped by how the interactant responds to the initial message, by how the emoter responds to the interactant's response, and so on. As Parkinson puts it, "[t]he acting out of emotion episodes is guided on-line by the affordances offered or denied by other people's ongoing actions, which in turn are mutually coordinated with the actor's own self-presentation" (163). This is a central idea in the theory of emotions as "urgency management systems" I develop in chapter 10.

Griffiths (2003) emphasizes that what is required to make sense of the idea that emotions are produced strategically is that appraisals take into account the likely impact on the interactant of engaging in an emotional episode. This is what Griffiths calls the *Machiavellian Emotion Hypothesis*:

> Emotional appraisal is sensitive to cues that predict the value to the emotional agent of responding to the situation with a particular emotion, as well as cues that indicate the significance of the stimulus situation to the agent independently of the agent's response (2003, 54).

Notice the two components of this hypothesis. On the one hand, appraisal is sensitive to the advantage of engaging in a certain emotion in certain circumstances. Under this view, the production of emotions is interactant-dependent, and contingent upon the effect expected from the emotional episode. On the other hand, appraisal is sensitive to cues that indicate the significance of the stimulus situation *as such*. Griffiths argues that the significance is independent of the *agents' response*, but I propose that we should understand it as independent of the *interactant's response*. In other words, emotional appraisal is sensitive to the significance of the antecedent circumstances for the emoter *both* independently of the interactant's response and in light of it.

I think the most fruitful way to interpret the Machiavellian Emotion Hypothesis is along the lines suggested by Hinde with respect to emotional expressions. I will reformulate it as follows:

> Emotional appraisal lies along a continuum from appraisal of interactant-independent personal meaning of antecedent circumstances to appraisal of interactant-dependent advantage of engaging in a certain emotion. Emotional appraisals sometimes lie close to the *pure registration of personal significance* end, and sometimes they lie close to the *pure negotiation of interpersonal advantage* end of the continuum. Most emotional appraisals, however, involve both aspects.

Consider an episode of anger. The appraisal that brings it about is sometimes close to the pure registration end of the continuum, for example when anger results from being suddenly poked in the back. In such case, the likely effect of getting angry on the interactant is not considered by the emoter, and anger unfolds automatically. One other occasions, the appraisal that brings anger about is close to the pure negotiation end of the continuum. For example, one may get angry with a store clerk just to get a refund on a purchased item, despite having lost the receipt and despite not blaiming the store clerk for it. In most cases, emotional appraisals involve both aspects, namely a registration of personal significance and a negotiation of interpersonal advantage.

The question we must ask is: What is the evidence that emotions are interpersonal moves, namely that they have a Machiavellian dimension? So far we have only relied on the intuitive appeal of this idea. Griffiths (2003) suggests that the evidence for the strategic effects of emotional displays in animals ought to be construed already as evidence that the production of emotion has strategic dimensions in animals. This is because the "distinction between having an emotion and expressing it may be distinctly problematic in nonhuman subjects" (2003, 56). Under this view, the strategic threat displays of a bird such as those discussed by Hinde must already be understood as evidence for the truth of the Machiavellian Emotion Hypothesis with respect to animals.

Griffiths also offers some tentative evidence about strategic dimensions of emotions in humans, which he argues ought to be expected on theoretical grounds given the homology between emotional appraisal systems in humans and animals. For example, Stein et al. (1993) reported that people list the prospect of obtaining compensation among the factors which determine whether they will appraise a loss as an anger elicitator or a sadness elicitor. Griffiths also mentions the evolutionary hypothesis that romantic love may have been selected for because of the differential fitness advantage afforded by mating outside one's long term bond (Konner

1982, 315). Although he acknowledges that there are competing explanations for the emergence of romantic love, Griffiths (2003) points out that Konner's provisional hypothesis is interestingly similar to the social constructionist interpretation of love as an excuse for adultery.

Parkinson et al. (2005) offer further evidence in support of the *emotions as interpersonal moves* thesis in humans. Consider embarrassment, an emotion elicited 98% of the time only in the presence of an interactant (Tangney 1996), and consequently a good candidate for the study of the transactional aspects of emotion. Parkinson reports that Leary, Landel, and Patton (1996) have shown that people continue to rate themselves as embarrassed about the way they sung a song ("Feelings") they just recorded until the experimenter somehow acknowledges that they are embarrassed. The consequence of the detection of embarrassment, as shown by Semin and Manstead (1982), is that interactants have a more lenient view of the embarrassing behavior.

Parkinson also reports that O'Malley and Greenberg (1983) showed that female experimental subjects reduced the fines they gave to motorists that expressed remorse in hypothetical accidents. This appears to offer us an insight into the effect of remorse on the interactor, which is likely to become a part of the appraisal that elicits it. There is also some preliminary evidence on the strategic elicitation of guilt (Miceli, 1992; Vangelisti, Daly, & Rudnick, 1991).

If we consider the evidence on the strategic dimensions of emotional expressions, and this preliminary evidence on Machiavellian effects in human emotions, the emotions as social transactions thesis appears worth taking seriously. I will try to develop its insights in the contaxt of a new theory of emotions as urgency management systems I will develop in chapter 10.

6.4. CONCLUSION

I have tried to disentangle various ways in which the thesis that emotions are social constructions has been understood. The most interesting strand of the social constructionist tradition has turned out to be the one which assimilates emotions with social transactions. As demonstrated by Griffiths' work in particular, this idea is compatible with affect program theory. As affect program theory, however, it appears to fall short as a general theory of emotions. There

clearly are instances of emotions which lack a strategic dimension. Reflex emotions such as fear generated by a suddenly looming object are an obvious example. But should a theory of emotions accommodate all instances of emotions as ordinarily understood? Isn't the point of a theory to help us formulate new explanations and predictions? In the next chapter, I begin tackling such methodological questions by illustrating and discussing the two most popular theories in contemporary philosophy of emotions, namely cognitivism and Neo-Jamesianism.

7. A CRITIQUE OF CONTEMPORARY PHILOSOPHY OF EMOTIONS

In contemporary philosophy of emotions, the two most popular theories are currently offered by cognitivists and Neo-Jamesians. Martha Nussbaum (2001) and Robert Solomon (2003), two prominent champions of contemporary cognitivism, have argued that emotions are particular kinds of judgments. Their theory is the most recent and sophisticated embodiment of the cognitivist tradition I characterized in chapter 4. Jesse Prinz (2004a, 2004b) has instead defended the view that emotions are particular kinds of perceptions of bodily changes. As I argued in subsection 2.3.1, this was the idea at the heart of James' (1884, 1990) theory of emotions. It has enjoyed a revival in recent times mostly through the work of the neurobiologist Antonio Damasio (1994, 1999, 2003). What Prinz adds to the Damasian version of James' theory is an account of the emotions' intentionality.

Cognitivists and Neo-Jamesians explicitly characterize their respective accounts of emotions as *definitions*, and claim that the sets of necessary and sufficient conditions they comprise capture anything which deserves to be called an emotion. When we probe the nature of this deservingness, we realize that their proposed definitions aim to capture anything that can rightfully be called an emotion in ordinary language. The implicit assumption is that the shared ground captured by the definition individuates a domain of theoretically interesting similarity among emotions as ordinarily understood. Although naturalistic philosophers of emotions such as Griffiths (1997) have voiced skepticism about the legitimacy of this project, I think it is an intellectually coherent project, which complements the project of developing explications of emotions.

My concern is of a different sort, namely that cognitivists and Neo-Jamesians presuppose that a satisfactory account of what makes something an emotion in the ordinary sense can take the form of a definition. There are good Wittgensteinian reasons to be skeptical about this assumption, and skepticism turns out to be well-grounded with respect to the work of

philosophers of emotions. When we consider their argumentative strategies in some detail, we realize that both cognitivists and Neo-Jamesians insulate their alleged definitions of emotions from ordinary language counterexamples by means of two equally problematic strategies. One is that of trivializing the notions used to define the emotions, and the other is that of legislating on proper use of ordinary language on the basis of introspective intuitions which are nothing more than expressions of prior theoretical commitments.

Neither of these strategies accomplishes either the goal of capturing the condition of application of folk emotion categories or the goal of developing a fruitful explication of them, which I take to be the two central purposes for a theory of emotions. If emotion theorists aim to accurately capture what makes something an emotion in the ordinary sense, then they should supplement their introspective intuitions with the empirical evidence collected by experimental psychologists in the last twenty years concerning the way in which ordinary language users sort emotions from non-emotions. This is what I will do in chapter 8, where I collect the empirical evidence on folk emotion categories and try to draw methodological conclusions from it.

On the other hand, if emotion theorists aim to develop fruitful explications of emotion categories, they should focus on demonstrating what theoretical purposes are served by the explications proposed, without striving for anything more than "similarity in use" with folk emotion categories. I discuss the properties a good explication of emotions should have in chapter 9.

7.1. WHAT ARE COGNITIVISTS AND NEO-JAMESIANS TRYING TO ACHIEVE?

Nussbaum and Solomon believe that "[e]motions are appraisals or value judgments" (Nussbaum 2001, 4) or that "to have an emotion is to hold a normative judgment about one's situation" (Solomon 2003, 8). Prinz (2004a, 55), on the other hand, is convinced that "emotions are perceptions of bodily states". These ostensibly very different accounts are understood as having the logical form of *definitions*. Consider the following passages:

> [W]e are in a position to conclude not only that judgments of the sort we have described are necessary constituent elements in the emotion, but also that they are sufficient (Nussbaum 2001, 43-44)
>
> Showing that [perceptions of] bodily changes are sufficient does not establish that the somatic theory is true. For that, one would also need to show that [perceptions of] bodily changes are necessary for emotions…I think the somatic approach can subsume anything that deserves to be called an emotion (Prinz 2004a, 46, 49)

Nussbaum (2001, 196) argues that emotions can "be defined in terms of judgment alone," and Prinz (2004b, 190) refers to "[m]y definition of emotions." But what makes a definition of emotions a good one? More generally, how do we establish whether or not a certain account of emotions, whatever its logical form, captures "anything that deserves to be called an emotion"? There are at least two paths a theorist of emotions may follow to articulate an answer.

One would be to say that a certain account captures anything that deserves to be called an emotion insofar as it captures what is worth calling an emotion relative to the purposes of a given theory. The key epistemic virtue that grounds deservingness is in such case *fruitfulness*. Under this view, good accounts of emotions would capture what deserves to be called an emotion *for the purposes of a theory T* (e.g. neuroscience, biology, experimental psychology, clinical psychology, social theory, etc.).

Another possibility would be to say that a certain account captures anything that deserves to be called an emotion insofar as it captures all and only what competent speakers can rightfully call an emotion in ordinary language. The key epistemic virtue that grounds deservingness would be in such case *ordinary language compatibility*.

If we consider the argumentative strategies of philosophers of emotions such as Solomon, Nussbaum and Prinz, it is apparent that the accounts they put forth aim for ordinary language compatibility (see below). At the same time, cognitivists and Neo-Jamesians assume that what deserves to be called an emotion in ordinary language corresponds to what is fruitful to call emotion for theoretical purposes, even though the specific nature of this fruitfulness is generally left implicit. The identification between what emotion terms allegedly mean in ordinary language and what is a "good thing to mean" by them for the purposes of a theory is one of the most common features of emotion theory writ large.[10] I am convinced that this is one of the central

[10] I borrowed the expression "good thing to mean" from Nuel Belnap.

methodological obstacles to progress in emotion theory. Accurately capturing the condition of application of folk emotion categories and offering fruitful explications for them are distinct intellectual projects, both interesting and legitimate, but to be evaluated in light of desiderata which are different and may not be fulfilled jointly.

I discussed James' theory of emotions in subsection 2.3.1 in terms of its content, but I now want to use it as an example of the way in which the epistemic virtues of theoretical fruitfulness and ordinary language compatibility have systematically been blurred in the history of emotion theory. It is useful to begin from James' theory also because cognitivism and the Neo-Jamesian theory of emotions are attempts to solve the substantive problems encountered by the Jamesian theory in the version I discussed in 2.3.1. Moreover, I want to argue that James would probably not have been a Neo-Jamesian, so it is good to refresh our memory about what James (1884, 1990) really said.

The inspiring thought of James' (1884, 1890, 1894) theory was that the psychology of emotions had yet to grow from a descriptive to a scientific phase. As I argued in chapter 2, psychologists had generally identified the emotions in terms of feelings, and treated such feelings as "eternal and sacred psychic entities, like the old immutable species in natural history." The "merely descriptive literature" resulting from the attempt to describe "the internal shadings of emotional feeling" was such that James declared he would rather "read verbal descriptions of the shapes of the rocks on a New Hampshire farm" than have to toil through it again (James 1890, 448).

What the psychology of emotions needed, James argued, was a "central point of view, or a deductive or generative principle," which could do for the understanding of emotions what the principle of heredity and variation had done for the understanding of species, namely allow it to get "on to another logical level" (James 1890, 448). Since James believed that feelings ultimately resulted from physiological causes, he proposed a physiological account of emotions.

When it came to spell out the *central reason* why his theory of emotions had to be embraced, however, James presented what I called the *Argument from Conceivability* (see 2.3.1 for discussion):

> I now proceed to urge the vital point of my whole theory, which is this: If we fancy some strong emotion, and then try to abstract from our consciousness of it all the feelings of its characteristic

> bodily symptoms, we find we have nothing left behind, no "mind-stuff" out of which the emotion can be constituted, and that a cold and neutral state of intellectual perception is all that remains… A purely disembodied human emotion is a nonentity…[F]or *us*, emotion dissociated from all bodily feeling is inconceivable (James 1884, 193)

Notice the shift of emphasis here. The "vital point" urged by James is not, as we may expect from the previous discussion, that there is significant heuristic value in thinking of emotions as perceptions of bodily changes, and that a large number of emotions, especially the strong ones, involve such perceptions. Rather, the vital point is that an emotion dissociated from all bodily feeling is "for us…inconceivable," in the sense that "whatever moods, affections, and passions" we can conceive of, they "are in very truth constituted by, and made up of, those bodily changes we ordinarily call their expression" (James 1890, 452).

James' argument presupposes the fruitfulness of thinking of emotions in terms of perceptions of bodily changes, but looks for legitimacy in the way in which competent language users – a.k.a. "we" – conceive of emotions. The result of this shift of emphasis has been that much of 20th century emotion theory has been focused on trying to show in how many ways not contemplated by James "we" can conceive of an emotion. This is not to say that conflicts between different schools have always been framed in terms of whether or not a certain account is compatible with ordinary language. For example, I argued in section 3.1 that psychological behaviorists questioned the Jamesian theory of emotions on the ground that it made use of a scientifically problematic notion, namely that of a *conscious* perception of bodily changes. The issue is rather that considerations of fruitfulness have been confusingly mixed with considerations of ordinary language compatibility.

Skinner (1953) had methodological misgivings about the notion of consciousness (see section 3.1.1), but he supported his theory of emotions in part by saying that "[w]hen the man in the street says that someone is afraid or angry or in love, he is generally talking about predispositions to act in certain ways" (1953, 162). The task of characterizing what predispositions the average person on the street is talking about when he or she talks about the emotions was of course perfected by Ryle (1949). I reported in section 3.2.1 on his distinction between the emotional predispositions he called *inclinations* (e.g. interest for symbolic logic), *agitations* (e.g. excitement) and *moods* (e.g. depression).

A formidable adversary for the Jamesian theory emerged when behaviorism collapsed around the mid-1950s, and a number of philosophers of emotions began working within the emerging cognitivist tradition along the lines I sketched in chapter 4. Solomon (2003) and Nussbaum (2001) are the contemporary champions of the cognitivist tradition, and they make abundant use of "conceivability" arguments. Consider the following quote from Nussbaum:

> The reason it makes sense to imagine a bodyless substance having genuine emotions is that it makes sense to imagine that a thinking being, whether realized in matter or not, could care deeply about something in the world, and have the thoughts and intentions associated with such attachments. And that's all we really require for emotion (Nussbaum 2001, 60)

What is important about this passage is that it clearly brings to the fore the nature of many conflicts between emotion theorists, which are conflicts about what "we really require for emotion." James' belief that a disembodied emotion is "for us" unconceivable is met by Nussbaum's assurance that "we" are perfectly able to conceive of a bodyless creature having an emotion. If we consider the sorts of ordinary language counterexamples discussed in debates between cognitivists and feeling theorists, and more generally in contemporary emotion theory, we realize that they belong to two families:

> *Type 1 Counterexamples*: Instances x of emotion (E) as ordinarily understood which do not satisfy the account A proposed
> *Type 2 Counterexamples*: Instances x which satisfy the account A proposed but are not instances of emotion (E) as ordinarily understood

In the passage above, Nussbaum offers a Type 1 counterexample to the thesis that an emotion is a perception of bodily changes. On the other hand, the Jamesian excerpt can be construed as a Type 2 counterexample to Nussbaum's thesis that an emotion is a judgment by which we "ascribe to things and persons outside the person's own control great importance for that person's own flourishing" (Nussbaum 2001, 4). As far as James is concerned, "we" would qualify such judgment as a "cold and neutral state of intellectual perception" insofar as it lacked bodily symptoms.

The prominence of Type 1 and Type 2 counterexamples in debates among emotion theorists reveals that the epistemic virtue of *ordinary language compatibility* is interpreted in terms of insulation from Type 1 and Type 2 counterexamples. Nussbaum explicitly declares that "[m]y procedure is Socratic," and she reminds us that candidate definitions are rejected in the Socratic dialogues by discovering either "what both Socrates and the interlocutor consider to be a genuine case…not covered by the definition" or by discovering that "the definition covers phenomena that neither Socrates nor the interlocutor is prepared to count as a genuine case" (2001, 10).

7.1.1. Counterexamples to cognitivism and Neo-Jamesianism

The attempt to formulate definitions of emotions capable of capturing what counts as an emotion in the folk sense ought to be met by Wittgenstein-style skepticism: Why are cognitivists and Neo-Jamesians convinced that "emotion" (or "anger," or "fear" etc.) are any different from "game" and innumerable other linguistic categories which cannot be defined? Didn't Socrates systematically fail at the task Nussbaum explicitly commits herself to? Prinz (2004a, 49) acknowledges that one "could concede that emotions form a mongrel category," and Nussbaum (2001, 8) points out that "some would claim that there is no interesting common ground" among the items that qualify as emotions in the ordinary sense. However, Prinz (2004a, 49) thinks that conceding that the emotions form a mongrel category "would leave us with a puzzle…Why does a single word, emotion, lord over such a motley?."

Ludwig Wittgenstein's (1953, § 66) well-known answer would be: "Don't say: 'There *must* be something common, or they would not be called [emotions]'- but *look and see* whether there is anything common to all." Cognitivists and Neo-Jamesians implicitly claim to have "looked and seen," and concluded that "the common ground within the class of emotions is actually greater than we might suppose" (Nussbaum 2001, 8) and that there is "far greater unity in the emotion category than often appreciated" (Prinz 2004a, 49).

I want to show that this claimed unity is achieved by means of two illegitimate argumentative strategies, namely the trivialization of the notions used to define the emotions, and the arbitrary stipulation that what fails to meet the proposed definition is not really an emotion. In this section, I clarify the nature of Neo-Jamesian and cognitivists definitions of emotions, and collect some of the most obvious Type 1 and Type 2 counterexamples they encounter. In the next section, I discuss the argumentative strategies employed by philosophers of emotions to insulate

their alleged definitions from such counterexamples, and argue that they fail. As it turns out, cognitivists and Neo-Jamesians have not followed Wittgenstein's advice, which is to "look and see" what emotions have in common. This task, which must be carried out empirically in order to be carried out reliably, is going to be tackled in the next chapter.

Let us begin from a quick summary of the main difficulties faced by the theory of emotions offered by James. As I mentioned, cognitivism and the Neo-Jamesian theory are alternative attempts to develop a theory of emotions capable of avoiding these difficulties. In his Argument from Conceivability, James had argued that "we" could not conceive of an emotion other than in terms of perception of bodily changes. He wrote about fear: "What kind of an emotion of fear would be left, if the feelings neither of quickened heart-beats nor of shallow breathing, neither of trembling lips nor of weakened limbs, neither of goose-flesh nor of visceral stirrings, were present, it is quite impossible to think" (James 1884, 183). Under this view, fear is the perception of a suite of bodily changes. This account appears to encounter Type 2 counterexamples, since something could be a perception of bodily changes of the kind described without being an emotion.

The first problem emphasized by early cognitivists in the 1950s and 1960s is that bodily changes are undifferentiated between distinct emotions. Bedford (1957) claimed, for example, that there is no difference between the bodily changes of indignation and annoyance, and he criticized James on this basis (see 4.1.3). Notably, this line of criticism has been a thorn in the side of the Jamesian theory since the physiologist Walter Cannon (1929, 352) argued that "[t]he responses in the viscera seem too uniform to offer a satisfactory means of distinguishing emotions which are very different in subjective quality."

The second and most serious problem is that the emotions appear to have intentionality. As I argued in chapter 4, we can distinguish two aspects to the intentionality of emotions. On the one hand, emotions are contingently about particular objects, at least most of the time. Emoters are generally afraid, angry or sad about particular individuals, events or states of affairs, rather than in an objectless fashion. On the other hand, emotions appear to be non-contingently about what Kenny first dubbed their *formal objects* (see discussion in 4.2.1). Subtleties apart, anger appears to be non-contingently about slights, fear about dangers, sadness about losses, and so on. This generates a dimension of normative assessment for emotions which has been called "logical", "conceptual" or "internal". The basic idea is that if the formal object of fear is the dangerous,

then fear is the sort of thing which is *inappropriate* when danger is not present. But "[i]f the emotions were internal impressions," Kenny (1963, 192) concluded, "there would be no logical restrictions on the type of object which each emotion could have."

If we interpret this argument as relying on the interpretation of formal objects as conceptually required descriptions of particular objects, then we run into trouble, because not all emotions are about particular objects. For example, anxiety and depression are often objectless. I argued in 4.2.1 that we should understand the formal object of an emotion as a description of its *conditions of satisfaction*. This is an alternative way to spell out the notion of intentionality for mental states, according to which a state X is intentional insofar as there is a way X is supposed to be (Searle 1983). Under this view, objectless anxiety is still intentional because it has conditions of satisfaction, namely the presence of danger.[11] Customary descriptions of formal objects for prototypical emotions are danger for fear, slight for anger, loss for sadness, transgression of a moral imperative for guilt, failure to live up to an ego ideal for shame, and others (see chapter 10 for an extended discussion).

Prinz's Neo-Jamesian theory contains interesting answers to the two criticisms I have reported. Firstly, Prinz has pointed out that the recent experimental literature on emotional physiology casts at least a reasonable doubt on the thesis that bodily changes are undifferentiated between different emotions (e.g. Levenson et al. 1990). Although this is true, I do not think that the evidence will ultimately born out a differentiation of emotions in terms of bodily signatures. This is because there is no theoretical reason why such differentiation should be expected. Emotions have several functions, and one of them is to prepare the body for action.

But insofar as different emotions are similar in terms of the actions to which they prepare, we should also expect that the bodily preparations they involve are similar. For example, there are similarities between what emoters do when they are ashamed and what they do when they are embarrassed (e.g. trying to avoid contact, trying to not to call further attention on themselves, etc.). Why should the bodily profiles of embarrassment and shame necessarily differ? What we ought to expect is only a difference between bodily profiles of emotions which predispose to very different actions. This may be the case for, say, fear and happiness. Experimental results

[11] I leave the discussion of whether all instances of objectless anxiety are inappropriate with respect to their formal object for another day.

appear to support the prediction that bodily differences exist at the level of families rather than at the level of individual emotions, with a few exceptions.

Cacioppo et al. (2000) published a meta-analysis of the available literature on physiological differentiation, which has so far focused mainly on anger, fear, happiness, disgust, and surprise. They reported than no study shows disgust to differ from control conditions on any measure of autonomic arousal. On the other hand, they reported several studies showing that heart rate increase is higher in fear than in anger, higher in anger than in happiness, higher in both fear and anger than in sadness, and higher in anger, fear, happiness, and sadness than in control conditions.

Also, diastolic blood pressure appears higher in anger than in fear, or sadness or happiness, and higher in sadness than in happiness. Anger also appears to differ from fear because it is associated with larger increases of nonspecific skin conductance responses, facial temperature, finger pulse volume, and smaller increases in stroke volume and cardiac output. Cacioppo et al. (2000) also suggested that there is a difference between the autonomic arousal of positive and negative emotions.

The distinction between positive and negative emotions is unclear, but for our purposes we can take positive emotions to be those manifested by appetitive approach and negative emotions to be those manifested by aversive avoidance. They argued that during negative emotions all autonomic indexes measured in the studies they surveyed (e.g. heart rate, diastolic pressure, blood volume etc.) were more active than during positive emotions. These results suggest that Cannon (1929) was both right and wrong (see 4.1.3 for background).

Cannon (1929) was right in thinking that emotions cannot be differentiated from one another (with possibly a handful of exceptions) in terms of the patterns of autonomic arousal eventually associated to them, because such patterns do not differ from one another to a sufficient degree. He was instead wrong in thinking that autonomic arousal is undifferentiated among different emotions. For example, very high increase in heart rate and diastolic blood pressure appears much more likely to be associated with anger or fear than with sadness.

But where does this leave the Neo-Jamesian theory? If Prinz's (2004a, 2004b) theory were that emotions differ from one another merely by virtue of bodily signatures, just saying that the evidence against their presence is inconclusive would not be much of a progress. But Prinz is not

committed to the thesis that the emotions must be differentiated from one another *exclusively* in terms of bodily changes.

Prinz's most intriguing innovation is having shown how perceptions of bodily changes could *appraise* in a way that at the same time is compatible with the Jamesian theory, contributes to type-identifying emotions potentially undifferentiated at the bodily level, and accounts for their intentionality. The key insight of Prinz's *embodied appraisal theory* is that emotions involve appraisals by virtue of what they represent, and that they represent what they have the function of being reliably caused by. This approach relies on Dretske's (1981, 1986, 1988) teleosemantic theory of representation, according to which, roughly speaking, mental states acquire conditions of satisfaction by virtue of being set up – by natural selection or by learning – to be set off by certain circumstances (Prinz 2004b, 54). Such circumstances consequently become those under which the state is the way it is supposed to be, namely those under which the state represents correctly.

Prinz's (2004a, 55) hypothesis is that "emotions are perceptions of bodily states…caused by changes in the body" and that such "changes in the body are reliably caused by the instantiation of core relational themes," from which it follows that the core relational themes are what the perceptions of bodily changes represent. In a nutshell, "[e]motions are states that appraise by registering bodily changes" (Prinz 2004b, 78), where bodily changes are assumed to have been set up to be set off by the emotion's core relational themes or formal objects. Fear is then not, as James would have it under a standard reading, *merely* the perception of a particular suite of bodily changes, but rather the perception of a suite of bodily changes *set up to be set off by danger*. Similarly, sadness is the perception of bodily changes set up to be set off by loss, shame is the perception of bodily changes set up to be set off by failure to live up to an ego idea, and so on. To generalize, the Neo-Jamesian definition of emotion E is that an emotion E is the perception of bodily changes set up to be set off by E's formal object.

Although this is certainly progress with respect to James' original account, the Neo-Jamesian theory still appears to encounter a large number of Type 1 counterexamples. Not only can we conceive of emotions as dispositions (e.g. love), but we can also conceive of emotions as not involving perceptions (e.g. unconscious anger) and of emotions as lacking bodily changes (e.g. guilt, shame). At first blush, what is true of emotion as a superordinate category is also true of particular emotions. We can conceive of fear as a disposition (e.g. fear of snakes), we can

conceive of fear as not involving bodily changes (e.g. fear that a certain politician will win the elections), and we can even conceive of forms of fear which do not involve a conscious experience of any kind (e.g. unconscious fear of failing).

Difficulties of this sort, and the conviction that feeling theories could neither handle the intentionality of emotions nor account for their type-identity, propelled the emergence of cognitivism in the 1960s and 1970s. In its most recent formulation, the cognitivist proposal is that "[e]motions are appraisals or value judgments" (Nussbaum 2001, 4). For example, Solomon (2003, 7, 8) argues that ""I am angry at John for taking…my car" *entails* that I believe that John has somehow wronged me… My anger *is* that judgment." Similarly, fear is the judgment that something dangerous is at hand, sadness is the judgment that a loss has been suffered, shame is the judgment that one has not lived up to an ego ideal, and so on. To generalize, the cognitivist definition of emotion E is that an emotion E is the judgment that E's formal object is instantiated.

The trouble with the cognitivist theory is that it also seems to encounter plenty of Type 1 counterexamples. For example, "we" conceive of instances of fear in the *absence of the capacity* for judgment (e.g. fear of cliffs in infants and animals), *prior to* judgment (e.g. reflex fears with a quick and automatic onset), and *in opposition to* judgment (e.g. fears elicited through a primitive appraisal mechanism in contrast with one's considered judgments as in spider phobias). Moreover, cognitivism appears to encounter Type 2 counterexamples, in the sense that even when emotions involve judgments, they do not seem to be *mere* judgments. Someone could judge that danger is at hand without being afraid, someone could judge that a loss has been suffered without being sad, someone could judge to have failed to live up to an ego ideal without experiencing shame, and so on.

Should we then conclude in the manner of Wittgenstein (1953, § 66) that "if you look at [the emotions] you will not see something that is common to *all*, but similarities, relationships, and a whole series of them at that"?

7.2. THE TROUBLE WITH COGNITIVIST AND NEO-JAMESIAN REBUTTALS

Cognitivists and Neo-Jamesians alike argue that we should not adopt a family resemblance account of emotions. As Prinz puts it, emotions "share a common essence. It is rare for nature (and folk psychology) to offer such a neat category." (Prinz 2004b, 102). Their argumentative strategy is to counter (a) Type 1 counterexamples by either claiming that their favorite definition is satisfied once we understand it properly, or that x is not really an instance of E in the ordinary sense, and (b) Type 2 counterexamples by either claiming that their favorite definition is not satisfied once we understand it properly, or that x is on reflection an instance of E in the ordinary sense. The way such strategy is carried out, however, comprises two problematic moves. One is to liberalize what is meant by respectively *judgment* and by *perception of bodily changes* by turning such notions into placeholders. The other is to let prior theoretical commitments color introspective intuitions about what counts as an emotion in the ordinary sense.

The Type 1 and Type 2 counterexamples I presented in the previous section have implicitly presupposed what we may call a *conservative* understanding of both *judgment* and *perception of bodily changes*. I have understood judgments along the lines of a philosophical tradition whose origin can be traced back to the Aristotelian and Stoic distinctions between an *impression* (phantasia) and a *judgment* (krisis). Under this view, judging is engaging in a mental operation of *assent* or *endorsement* with respect to some propositional object p.

In more recent times, Sellars (1966, 150) has offered a broadly Kantian theory of judgment according to which judging that p is engaging in a *discursive inner episode*, which has as its model the *overt verbal reporting* that p (see Brandom (1994, 2000) for a sophisticated development of this idea). According to this understanding of judgment, the activity of judging that p is the paradigmatic expression of linguistic abilities, it is reflective, it is sensitive to the evidence that p, it requires possession of the concepts deployed in p, it requires the ability to recombine such concepts in the context of other propositional contents, and so on. Similarly, I have understood the notion of *perception of bodily changes* in terms of the notion of conscious perceptual experience of autonomic changes. It is under this conservative understanding of *judgments* and *perception of bodily changes* that the counterexamples I presented in the previous section have sounded persuasive.

7.2.1. The Placeholder Strategy

Cognitivists and Neo-Jamesians have responded to their critics by arguing that this is not what they *meant* by, respectively, judgments and perceptions of bodily changes. Let us consider the cognitivist rebuttals first. Since we can conceive of infants and animals having emotions, but judgments understood conservatively are not available to them, the cognitivists have simply stated that infants and animals are capable of assenting to propositional objects. As Solomon (2003, 187) puts it with possibly ironic disregard for well-known philosophical controversies, "I take it as uncontroversial that animals make all sorts of judgments." Since emotions can be unconscious and can be elicited through primitive appraisals, whereas judgments understood conservatively are deliberate and involve higher cognitive abilities, cognitivists have allowed judgments to be instantiated at low levels of cognitive complexity. For example, Solomon describes judgments as not being necessarily "conscious or deliberative or even articulate," and says that sometimes "we judge unconsciously, without thinking or reflection" (Solomon 2003, 210-211).

Since emotions sometimes involve physiological changes, whereas judgments understood conservatively simply amount to ways to make up one's mind, the cognitivists have assumed that judgments comprise changes in the body. Nussbaum (2001, 45) argues that "I am conceiving of judging as dynamic, not static…So why would such a dynamic faculty be unable to house, as well, the disorderly motions of grief?." Solomon makes the point even more explicitly, as he says that "[o]ne can, and sometimes must, speak of bodily judgments" (Solomon 2003, 213). Since emotions involve behaviors and behavioral tendencies, whereas judgments understood conservatively are not causally efficacious in the absence of a conative attitude, Solomon argues that "judgments in emotions are judgments which have a quasi conceptual connection with desires," in the sense that "one might analyze the various emotions as judgmental structures enclosing a core desire which is both their motivation and their "conatus"" (Solomon 2003, 105-106). Since emotions involve mental behaviors such as the recruiting of memories, Nussbaum (2001, 65) states that "[g]rief is not just an abstract judgment [in the sense that] the experience itself involves a storm of memories."

There are several other examples of this liberalizing strategy at work, but I take it that its general thrust is already clear. Judgment is slowly turned into a placeholder, which on the one hand comprises the actual set of properties something must fulfill in order to be an emotion in

the ordinary sense (e.g. the property of including a behavioral tendency), and on the other hand allows for something to qualify as an emotion in a variety of different ways (e.g. with and without physiological changes, consciously or unconsciously, etc.).

The same unbridled liberalizing strategy characterizes Prinz's Neo-Jamesian rebuttals. Since some emotions are unconscious, and this appears to be in contrast with James' thesis that all emotions amount to perceptions, Prinz has argued that perceptions of bodily changes need not be conscious. In this respect, he is following the neurobiologist Damasio, who has argued that "a signal body state or its surrogate may have been activated but not been made the focus of attention. Without attention, neither will be part of consciousness, although either can, be part of a covert action on the mechanisms that govern, without willful control, our appetitive (approach) or aversive (withdrawal) attitudes toward the world" (Damasio 1994, 190). Since some emotions appear to lack bodily changes of the autonomic variety, whereas James had claimed that all emotions are perceptions of bodily changes, Prinz expands the notion of bodily change to comprise neural counterparts of autonomic changes.

This is another suggestion borrowed from Damasio, who argued that there are two types of neural bodily changes, those generated by the "body loop" and those generated by the "as if body loop" (Damasio 1999, 281). Both are representations of the "body landscape" in "somatosensory structures of the central nervous system," but whereas "body loop" changes correspond to actual changes in the body, "as if body loop" changes do not correspond to any changes in the body other than neural ones. In the "as if" case, "the representation of body-related changes is created directly in sensory body maps, under the control of other neural sites, for instance, in the prefrontal cortices" (Damasio 1999, 281).

In some passages, Prinz appears to liberalize the notion of bodily changes even further, to encompass also the instrumental and expressive behaviors involved in the emotions. For example, he argues that "a somatic change can be a change of facial expression, an increase in heart rate, a secretion of hormones," and reports approvingly that James considered "bodily changes" to comprise "everything from tremors and tears to striking out in rage" (Prinz 2004b, 5). It seems to me that a notion of *perception of bodily changes* as liberal as the one recommended by Prinz is in conflict with the basic tenets of the Jamesian theory. James described his theory as a "physiological theory" of emotions, and justified this label by arguing that "the general causes of the emotions are indubitably physiological" (James 1890, 449), a

qualification which suggests that the bodily changes whose perception is the emotion are to be understood as autonomic bodily changes.

The spirit of the Jamesian approach is arguably still maintained when autonomic changes are turned into their neural counterparts. But it is certainly lost when instrumental behaviors and expressions can count as bodily changes, and when perceptions of them can become unconscious. Unconscious aggressive behavior towards one's mother, deprived of either autonomic changes or of their neural counterparts, may qualify as a *perception of bodily changes* if we understand such notions as liberally as Prinz appears willing to, but it is certainly not the sort of thing James meant when he said that "every one of the bodily changes, whatsoever it be, is FELT, acutely or obscurely, the moment it occurs" (James 1884, 192).

Notice that in the crucial passage in which James tries to "to urge the vital point of my whole theory," the bodily manifestations mentioned are *not* behaviors (expressions or instrumental behaviors), but rather autonomic responses. James (1884, 193) did not ask: What would be left of fear if we didn't run, or if we didn't display the facial expression of fear, but, rather, what would be left of fear without having a feeling, i.e. a conscious perception, of "quickened heart-beats," "shallow breathing," "trembling lips," "weakened limbs," "goose-flesh," and "visceral stirrings." This suggests that, even though James in some passages did refer to instrumental behaviors as "bodily manifestations," this is not what he meant by bodily changes in the context of his trademark thesis that "*our feeling of [bodily] changes as they occur* IS *the emotion"* (James 1884, 189-190).

My remarks are meant to question the fittingness of the *Neo-Jamesian* label when it comes to designating Prinz's and Damasio's theories, because the notion of bodily changes they endorse strikes me as unfaithful to the spirit of James' original enterprise. When a Freudian unconscious anger towards one's mother or a Rylean interest in symbolic logic can both count as special cases of the Neo-Jamesian theory, it is legitimate to ask what the label Neo-Jamesian is supposed to designate.

The real issue, however, is not exegetical. The real issue is that Prinz is in effect doing to the notion of *perception of bodily changes* what cognitivists have done to the notion of *judgment*, namely turn it into a placeholder. To say that striking out is *the bodily change whose perception is rage* is the strategic equivalent of saying, as proposed by Solomon, that striking out is the result of the desire for revenge included quasi-conceptually in *the judgment that one has been*

slighted which is rage. Both are highly misleading ways to say that rage involves not only an appraisal of slight and a perception of physiological changes (which may not be necessary), but also a behavioral tendency of attack expressed by striking out.

The main problem with turning an alleged *definiens* of emotions into a placeholder is that this procedure hides from view precisely what needs to be clarified. When we try to find out what makes something an emotion/rage in the ordinary sense, we are not looking for a generic label to designate the condition whose fulfillment makes something an emotion/rage. We are looking for that very condition, in all of its complexity.

The strategy of liberalizing what is meant by *judgment* and by *perception of bodily changes* to the limit of their meaning in English allows maintaining the impression that one has defined the emotions, but it is in effect a way to avoid asking what the members of folk emotion categories share by virtue of which they are members.

The alternative, of course, is to explicate emotions, without trying to capture conditions of application for folk emotion categories. This is ultimately the strategy I will recommend, but it is not the strategy philosophers of emotions are following. The nature of the debate in which they are engaged clearly reveals that they are interested, among other things, in achieving ordinary language compatibility.

7.2.2. Legislating on ordinary language

There is a second, and equally problematic, strategy adopted by cognitivists and Neo-Jamesians to defend their alleged definitions of emotions, namely to legislate on proper use of ordinary language. In this case, instead of liberalizing what is meant by either *judgment* or *perception of bodily changes*, it is stipulated that what fails to satisfy the definition proposed is not a real emotion. For example, Solomon asks: “Do we make any judgment at all when we are simply startled? I suspect the answer is no” (Solomon 2003, 214). Solomon does not explain why we do not, and I will not try either, although I do not see any reason why we should make a solitary exception for startle given that judgment plays in effect a placeholder function within the cognitivist theory.

The interesting part is that Solomon (2003, 214) argues that “there is considerable dispute whether the startle reaction is an emotion at all, and that question is firmly focused on the question of whether it involves a judgment.” This provision guarantees that the cognitivist

account will accommodate startle, because the only way for startle to *really* be an emotion is being a judgment. But this is nothing other than the expression of a theoretical commitment. No evidence is provided that startle is or is not an emotion as ordinarily understood, and this certainly cannot be established by asking whether or not startle is a judgment. In the case of startle, Solomon is guaranteeing the appropriateness of cognitivism by mere stipulation.

Prinz (2004a, 50) argues with respect to the aesthetic emotion of delight experienced by an art critic contemplating a piece of artwork that "[i]f the critic claimed to find delight in an artwork but showed absolutely no somatic response, we might justifiably question her sincerity." In this case as well, it is quite surprising that Prinz even contemplates this possibility, since the notion of somatic response has been expanded so as to include instrumental behaviors which are clearly involved in the critic's delight (e.g. approaching, looking at the artwork intently). In the passage cited, Prinz seems to be resorting to a *conservative* understanding of somatic responses as autonomic responses, along the lines I have claimed James presupposes all along.

The interesting point is that Prinz argues that in order to *really* be an emotion, delight needs to involve a somatic response. No evidence is offered that delight ordinarily counts as an emotion just in case it involves autonomic changes. It is simply stipulated that "we might justifiably question" the critic's sincerity if it did not. Prinz inherits this argumentative strategy directly from James (1884), who also discussed the case of what he called *subtler emotions* such as aesthetic, moral and intellectual ones. James argued that in the case of subtler emotions "[t]he bodily sounding-board is at work, as careful introspection will show, far more than we usually suppose," but he concluded that in all cases in which it is not, "a cold and neutral state of intellectual perception [would be] all that remains" (James 1884, 201).

What is problematic with this second strategy is that the theorist is trying to have it both ways, namely capturing what counts as an emotion in the ordinary sense and stipulating without supporting evidence that what does not satisfy the proposed definition is not an emotion in the ordinary sense. The normative commitments of the theorist in effect determine what things count as emotions as ordinarily understood. This shortcoming is grounded in the exclusive reliance of cognitivists and Neo-Jamesians on their own introspective intuitions when it comes to establishing what things counts as emotions as ordinarily understood. I do not mean to be saying that philosophers of emotions make no use of data other than own introspective intuitions. Prinz (2004b, 30) explicitly argues that "philosophical methods are most powerful when used in

conjunction with empirical data," and Nussbaum (2001, 9) asserts that in her account "intuitive judgments about…cases are consulted throughout, along with the results of…scientific investigations." Both Neo-Jamesian and cognitivist philosophers of emotions are surely up to date on the latest scientific discoveries *about the emotions*.

What I am saying is that no empirical data are considered concerning the way in which ordinary language users classify instances of emotion, or of particular emotions. The introspective intuitions of the theorist about what items count as emotions are acknowledged to be fallible, and provisions are made for the possibility that scientific discoveries down the line will show "our" ordinary conceptions of emotions to be at least in part mistaken (Prinz 2004b, 29; Nussbaum 2001, 9-10). The point is that the theorist's intuitions are the only evidence considered by Neo-Jamesians and cognitivists concerning what "we" currently conceive an emotion to be.

This strikes me as a major source of unproductive disputes in emotion theory, in which two authors respectively affirm and deny on the basis of their clashing introspective intuitions that "the man in the street" means X when he or she conceives of emotions. My view is that "relying on people's ability to classify instances of emotion," as Nussbaum (2001, 9) suggests we should, requires studying empirically how such ability gets to be manifested across the language community.

Nussbaum formulates an analogy which strikes me as a helpful way to describe what many philosophers of emotions are in effect trying to achieve. Nussbaum (2001, 9-10) compares the job of an emotion theorist to that of a field linguist who is trying to uncover rules of grammar by relying on judgments of grammaticality. Such judgments are not accompanied by the ordinary speakers' ability to tell what makes a sentence grammatically correct, but what the grammarian is trying to describe are the rules of correctness tacitly embedded in ordinary judgments of grammaticality.

Clearly, some language users will make mistakes about what they take to be grammatical. But the job of the theorist is to extract from the total set of ordinary judgments of grammaticality, the correct and the incorrect ones alike, the rules such as they may exist concerning what makes a sentence grammatical. Cognitivists and Neo-Jamesians are trying to do the same thing with respect to emotion categories. They aim to extract from ordinary classifications of items as emotions, correct and incorrect ones alike, what the condition of application for emotion kind

terms are in ordinary language. The point is that if this is what they are trying to achieve, they need to ground their theories in careful field work. As the history of emotion theory vividly shows, the intuitions of theorists clash, and they do so partly because they often do not express anything other than prior theoretical commitments. Fortunately, a significant amount of field work has already been done by experimental psychologists in the last twenty years. In the next chapter, I will summarize what they have discovered.

7.3. CONCLUSION

The history of emotion theory is a long sequence of attempts to individuate a subset of marks of emotionality such that anything that deserves to be called an emotion fulfills them. I pointed out that this notion of deservingness has generally been ambiguous between two interpretations, namely that of *theoretical fruitfulness* and that of *ordinary language compatibility*. Emotion theorists have been unclear about what sort of deservingness they were pursuing, in part because of the widespread assumption that what emotion terms "mean" coincides with what is a "good thing to mean" by them relative to the purposes of a theory.

Cognitivists and Neo-Jamesians are the latest representatives of the long lineage of theorists working under this assumption. As revealed by their argumentative strategies, they strive to come up with definitions of emotions which are ordinary language compatible, and assume that such definitions capture a theoretically interesting common ground between emotions. Cognitivists and Neo-Jamesians are aware of the fact that most lexical categories cannot be defined in terms of necessary and sufficient conditions.

However, they seem to be convinced that, although "it is rare for nature (and folk psychology) to offer…a neat category" (Prinz 2004b, 102), emotion categories just happen to be such rare categories. I have tried to show that that the alleged "neatness" of emotion categories is entirely illusory, because it results from the equally problematic strategies of (a) trivializing the ingredients used in the definition to the limit of their meaning in English and (b) stipulating that the only way for something to really be an emotion is to fulfill the proposed definition.

My worries about cognitivism and Neo-Jamesianism, I emphasize it, go deeper than this. As I will argue in chapter 10, these theories fail to offer a viable account of the intentionality of emotions, and disregard the fundamental motivational dimension they have. What I focused on in this chapter are the sorts of problems cognitivism and neo-Jamesianism encountered achieving what they set out to achieve, namely offering a definition of emotions which does not encounter ordinary language counterexamples.

8. INTERPRETING THE EMPIRICAL EVIDENCE ON EMOTION CONCEPTS

In the first seven chapters of this dissertation, I explored some of the most influential theories of the emotions ever proposed, studying a number of influential figures from Ancient Greece to our time. At first blush at least, every account I considered seems to encounter a domain of emotional phenomena to which it fails to apply, unless ad hoc moves are put into place. I have just explored in some detail the shape this problem takes with respect to cognitivism and Neo-Jamesianism, but I emphasize that it is a general problem.

An affect program theorist such as Ekman, for example, claims that all emotions are basic emotions, which, among other properties, are assumed to be short-lived episodes associated with typical facial expressions (see subsection 5.1.3). But very many emotions appear to be long-lived episodes, and lack such expressions. For example, guilt is generally not associated with a guilt-typical facial expression, and it generally lasts for significant periods of time. Ekman's response to these sorts of counterexamples is in effect to legislate on ordinary language. His point is simple: "I do not allow for "non-basic" emotions". Under this view, guilt is not an emotion. But this position raises the need for a rationale, which can't be found in Ekman's own theory. Are we free to legislate however we want when it comes to giving a theory of emotions?

Social constructionist theories of emotion also encounter domains of emotional phenomena they are unable to account for. As I argued in subsection 6.3.2, Averill proposes that emotions are transitory social roles interpreted as passions. But some emotions do not fit this model, as there doesn't seem to be anything social about them. For example, it is hard to understand how the sort of fear one experiences when suddenly losing support may amount to a social role. One could of course start twiddling with the notion of a social role so as to make it fit every case, but ad hoc moves such as these are not to be recommended, because their explanatory payoff is nil

It is time to understand why emotion theorists have had so much trouble capturing the emotional domain with definitions which apply to them all, and with generalizations that admit

of no exceptions. The way to do it is in my view to heed the advice Wittgenstein gave about games, namely to "look and see" what sorts of things emotions have in common by virtue of which they are called emotions.

In the past twenty years, psychologists have empirically studied emotion concepts, by which they refer to the kinds of mental representations that govern people's categorizations and inferences with respect to emotion categories. Categories named by emotion terms such as "emotion", "fear", "anger" etc. are in turn assumed to be sets of items which fulfill a condition of category membership, which describes the condition of application for the emotion term which designates the category.[12]

What this (generally neglected) empirical literature reveals is that, as we may expect, emotion categories manifest prototypicality phenomena, since some emotions are judged to be better examples of "emotion" than others. But the literature also suggests that emotion categories manifest a great deal of heterogeneity and vagueness. Heterogeneity and vagueness present emotion theorists with a puzzle, which is how one should go about studying items of a category that contains instances which are widely different from one another as well as borderline instances.

This is the puzzle I try to solve in chapter 8, where I explain what I take to be the desiderata for a good theory of emotions. In this chapter, I illustrate a number of key empirical facts about emotion concepts, and offer an interpretation of what they tell us about folk emotion categories.

8.1. EMOTIONS AND PROTOTYPICALITY

The first systematic empirical study of the nature of emotion concepts is due to Fehr and Russell (1984). They began their study by asking 200 experimental subjects – undergraduates from the University of British Columbia - to freely list examples of members of the category of emotion (generation of categories subordinate to emotion), and to freely list the general category of which categories such as anger, love, fear, sadness were instances (generation of categories

[12] I will use the terms "category" and "kind" interchangeably, under the interpretation I just mentioned.

superordinate to anger, love, fear etc.). This experiment was meant to study the *hierarchical structure* of emotion categories. In the first experiment, 196 examples of emotion were listed by at least two subjects.

The following are the ten items freely listed by most subjects as instances of the category "emotion": Happiness (152/200), Anger (149/200), Sadness (136/200), Love (124/200), Fear (96/200), Hate (89/200), Joy (82/200), Excitement (53/200), Anxiety (50/200), Depression (42/200). The items at the bottom of the ranking are things like Tranquility, Ambivalence, Withdrawn, Weak, Wanting, Uptight, Unstable, Understanding (each freely listed by no more than 4 people). More than half the subjects freely listed happiness, anger, sadness and love as instances of "emotion".

In a second study, Fehr and Russell (1984) chose ten items other than the first ten from the list obtained in the first study, looking for instances of emotion cited at different levels of frequency. The ten items selected were the following: Disgust, Guilt, Embarrassment, Worry, Awe, Pride, Envy, Calmness, Boredom, and Respect. A list was then formed by combining such selected items with the 10 most frequently listed categories of emotion.

Fehr and Russell (1984) called the set of 20 items so obtained "target emotions". The target emotions were then distributed in four 20-item lists, together with filler items such as tingle, stubbornness, and alertness. The experimental subjects were asked to provide a superordinate category for each of the items in the four lists. "Emotion" was free listed as a superordinate category for most of the target emotions, but with widely different frequencies. For example, whereas 64.3% of the subjects indicated emotion as a superordinate for "love", only 3.3% indicated emotion as a superordinate for "boredom", and 0% indicated emotion as a superordinate for "respect".

These results suggest that "emotion" represents the *head of a hierarchy of categories,* and that the subordinate-superordinate relation between "emotion" and specific emotions comes more quickly to mind for some emotions rather than for others. The results also suggest that there may be some disagreement in the population of speakers concerning whether or not something counts as an emotion, an issue we will explore in the next section.

One of the key results presented by Fehr and Russell (1984) is that "emotion" instantiates protypicality effects, namely that some of its members are judged to be better examples and some of its members are judged to be worse examples of emotion. The phenomenon of

prototypicality is very widespread, so it would be highly surprising if it did not apply to emotions as well. Evidence of prototypicality effects has been shown with respect to biological categories (Rips, Shoben and Smith 1973), trait and person categories (Cantor & Mischel, 1977), social psychological categories (Cantor, Mischel, & Schwartz, 1982), clinical categories (Cantor, Smith, French, & Mezzich, 1980), categories of painting style (Hartley & Homa, 1981), categories of musical themes (Welker, 1982), and even categories such as "even number" or "female" (Armstrong et al. 1983).

The importance of the phenomenon is that "the prototypicality of items within a category can be shown to affect virtually all of the major independent variables used as measure in psychological research" (Rosch 1978, 198). For example, judgments of prototypicality have been proven to predict reaction times in sentence verification tasks (Hampton, 1979; Rosch and Mervis, 1975), order of output when asked to provide an instance of the kind (Barsalou & Sewell, 1985), efficacy in priming tasks (Rosch 1975; Rosch, Simpson, & Miller, 1976), order in which they are learned by infants (Rosch 1973), and drawing of inferences from being a category member to having a certain property characteristic of the category (Smith 1989).

A variety of explanations have been proposed for judgments of prototypicality with respect to a member x of a category, most importantly (a) how similar x is to a mentally stored exemplar of the category or to an abstraction of central tendency of the category, (b) how close x is to an ideal of the category, (c) how frequent is the instantiation of x within the category (see Barsalou and Sewell 1985).

Rosch and Mervis (1975) offered empirical evidence that "members of a category come to be viewed as prototypical of the category as a whole in proportion to the extent to which they bear a family resemblance to (have attributes which overlap those of) other members of the category. Conversely, items viewed as most prototypical of one category will be those with least family resemblance to or membership in other categories" (Rosch and Mervis 1975, 575).

Notice that there may be family resemblance in this sense whatever the condition of membership for the category is. For example, if most members of the kind *grandmother* share, besides the condition of membership of being the mother of someone's mother or father, the properties of having white hair, moving with some difficulty, and making presents at Christmas, then grandmothers with these further properties will be judged more typical of the category than

grandmothers without such properties, as they "have attributes which overlap those of other members of the category".

At the same time, if the very condition of membership to the category were family resemblance, prototypicality phenomena would ensue whether or not there are differences with respect to the overlap of attributes besides those that make them kind members. Kind members may already be viewed as prototypical to the extent to which they have cluster properties in common with other members of the category.

A category such as "tall" may show prototypicality effects for a different reason, namely that the best example of a tall man is that of a man as tall as possible, namely as close to an ideal of tallness as possible.

Finally, a category such as "even number" may show prototypicality effects because the number 4 is more frequently instantiated in the life of cognizers than the number 106. These elements can of course combine to determine judgments of prototypicality (see Murphy 2003 for review). The fact that also categories such as "even number" show prototypicality effects, incidentally, indicates that prototypicality effects *as such* do not count as evidence that the condition of membership of a category has any particular logical form rather than another.

We can now ask: What are the best examples of "emotion" according to competent English speakers? In a third study, Fehr and Russell (1984) asked subjects to evaluate how good an example of emotion each of their 20 "target emotions" was. They graded the judgment of prototypicality by asking subjects to assign to each emotion a number of points ranging from 1 (for an "extremely poor example") to 6 (for an "extremely good example"). The following is the ranking of prototypicality they obtained for the twenty target emotions:

Average prototypicality ratings for 20 "target emotions"
1. Love (5.46/6.00)
2. Hate (5.26/6.00)
3. Anger (5.15/6.00)
4. Sadness (5.04/6.00)
5. Happiness (5.00/6.00)
6. Joy (4.89/6.00)

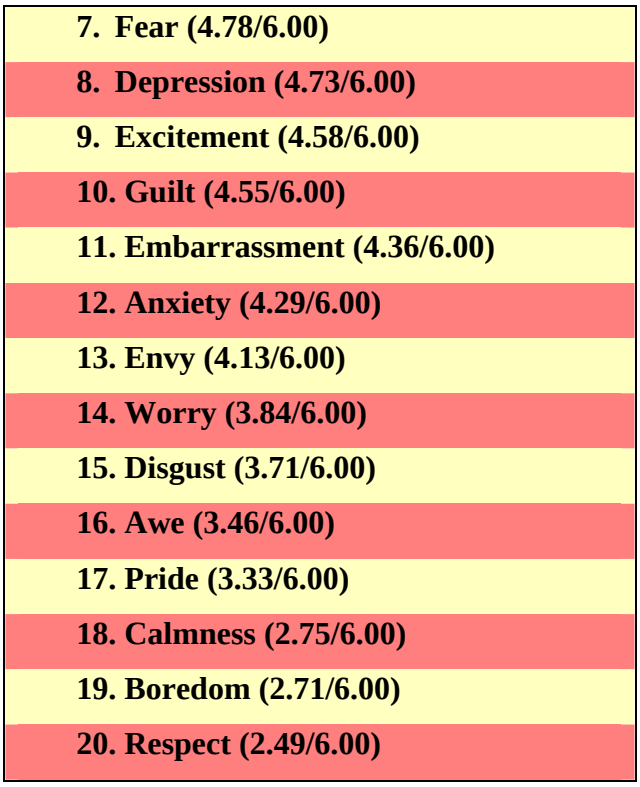


Figure 7: Which emotions are prototypical? From Fehr and Russell (1984)

The data indicate that, among the 20 "target emotions", "love", "hate", "anger", "sadness" and "happiness" are the best examples of emotion, and "calmness", "boredom" and "respect" are the worst examples. There are two limitations to this experiment. The first is that it does not give us a complete overview of prototypicality phenomena for the category of "emotion", because it is limited to 20 target emotions, 10 of which were chosen among items freely listed by just a handful of people in the first study (e.g. awe was free listed by 4 people only). The second is that Fehr and Russell's (1984) choice scheme did not allow subjects to distinguish between the judgment that a target emotion is an extremely poor example of emotion and the judgment that it is not an emotion at all. This means that judgments at the bottom of the scale are ambiguous between judgments of low prototypicality and judgments of non-membership.

Be that as it may, Fehr and Russell (1984) demonstrated that the prototypicality of the twenty target emotions was correlated with a number of indexes of cognitive performance, including frequency in a free-listing task, probability that "emotion" was mentioned as a superordinate category when prompted and substitutability for the term "emotion" in natural sounding sentences while maintaining their "naturalness." For example, Fehr and Russell (1984,

472-474) demonstrated that sentences which "sound natural" about emotion – presumably because they express received ideas about the category - start sounding unnatural once "emotion" is substituted by non-prototypical emotions. On the other hand, the sentences maintain their natural sounding quality if the substitution is made with prototypical emotions. Just to give a couple of examples, "emotion enables the individual to exert great energy for a brief period" and "emotion means an aroused or "stirred up" frame of mind" sound natural when "anger" or "fear" are substituted to "emotion," but not when "remorse" and "melancholy" are (Fehr and Russell 1984, 472-474).

This suggests that the received ideas people have about emotions, as about any members of a category with prototypicality effects, tend to be formed around prototypical instances, namely those which are considered to be good examples of the category, come to mind first, are learned earlier, and so on.

8.2. EMOTIONS AND VAGUENESS

I suggested before that Fehr and Russell's (1984) data on prototypicality ratings fail to distinguish between poor examples of emotion and non-instances of emotion. Fortunately, their article contains studies which allow us to investigate this distinction. Fehr and Russell (1984) asked 37 subjects whether each of the 20 "target emotions" they had selected was indeed an emotion. Their sample was not very large, but I will assume for the sake of argument that the result is representative of the linguistic community as a whole. This assumption is supported by the fact that introspection leads to data compatible with those offered in this experiment. Here is a summary of their main result:

"Target emotion"	**Percentage of subjects who said it is an emotion**	**Percentage of subjects who said it is not an emotion**
Love (5.46/6.00)	**94%**	**6%**
Hate (5.26/6.00)	**100%**	**0%**

Anger (5.15/6.00)	**100%**	**0%**
Sadness (5.04/6.00)	**100%**	**0%**
Happiness (5.00/6.00)	**100%**	**0%**
Joy (4.89/6.00)	**96%**	**4%**
Fear (4.78/6.00)	**97%**	**3%**
Depression (4.73/6.00)	**89%**	**11%**
Excitement (4.58/6.00)	**92%**	**8%**
Guilt (4.55/6.00)	**89%**	**11%**
Embarrassment (4.36/6.00)	**78%**	**22%**
Anxiety (4.29/6.00)	**82%**	**8%**
Envy (4.13/6.00)	**84%**	**16%**
Worry (3.84/6.00)	**89%**	**11%**
Disgust (3.71/6.00)	**94%**	**6%**
Awe (3.46/6.00)	**74%**	**26%**
Pride (3.33/6.00)	**74%**	**26%**
Calmness (2.75/6.00)	**52%**	**48%**
Boredom (2.71/6.00)	**61%**	**39%**
Respect (2.49/6.00)	**26%**	**74%**

Figure 8: Which emotions are borderline? From Fehr and Russell (1984)

Although in some cases the presence of a small percentage of subjects who disagree from the majority is to be understood as the equivalent of mistaken judgments of grammaticality, this interpretation becomes less convincing the closer the population of experimental subjects becomes to being equally divided. For example, whereas we can think of the 3% and 4% of experimental subjects who believe that, respectively, fear and joy are not emotions as being incompetent language users, this interpretation is distinctively less appealing when roughly 75% of the people have a certain view about membership and one quarter disagree (e.g. with respect to pride, respect, awe, embarrassment). In cases such as *calmness* (52% vs. 48%) and *boredom* (61% vs. 39%), the experimental subjects are fairly equally divided on the emotion status of the category.

I call these examples of "epistemic indeterminacy" with respect to category membership, namely cases in which a significant amount of language users are in disagreement as to whether or not some x definitely belongs to some category C. There are other possible manifestations of "epistemic indeterminacy", although Fehr and Russell (1984) did not study them with respect to emotion. For example, sometimes a significant amount of language users are uncertain, rather than disagree, about whether or not some x definitely belongs to the category, or think that x belongs to the category only to a certain degree. This appears to be the case for example for categories such as "bald" or "tall". Language users appear to be systematically uncertain as to whether or not certain people are tall or bald, and/or state that they are tall or bald only to some degree.

But what is the explanation of "epistemic indeterminacy"? The answer, it seems to me, is to be given on a case by case basis. On some occasions, the best explanation of it will be ignorance of the condition of membership on the part of language users, or inability to verify whether or not a known condition is fulfilled because of a cognitive failure of some kind. For example, a significant amount of language users are likely to be in disagreement or uncertain as to whether or not the square root of 16 is even. Some may not know the condition of membership for "even", i.e. being a natural number exactly divisible by two, and some may be unable to calculate the square root of 16.

In some cases, however, there seem to be no evidential reasons to conclude that epistemic indeterminacy results from ignorance or cognitive deficit on the part of language users. An alternative explanation for epistemic indeterminacy is *vagueness*, which I understand in the manner of non-epistemicist theories of vagueness as lack of sharp boundaries for the category. As Sorensen (2003) puts it, under this view of vagueness "[b]orderline cases are inquiry resistant", in the sense that there simply is no fact of the matter as to whether they fall under the category. Our habits of language have left it indeterminate whether or not borderline cases belong to the category. The alternative epistemic interpretation of vagueness, championed most prominently by Williamson (1994), takes ignorance to be the appropriate explanation for borderline cases. Categories are assumed to be always such that their condition of membership settles whether or not an item belongs to them, although often language users ignore what such condition is.

To defend the view of vagueness as lack of sharp boundaried from objections, and to show it to be superior to non-epistemic theories of vagueness, is a complex job I cannot undertake in this dissertation. I refer the reader to existing defenses of non-epistemic theories of vagueness (e.g. Wright (2001), see Sorensen (2003) for further references), and assume henceforth that a vague category is one whose condition of membership does not settle for all x whether or not they definitely belong or fail to belong to the category.

The category "emotion" is in my view very much like categories such "religion", "game", "democrat", which are what Alston (1967) called "combinatory vague" categories. They are vague because, as Wittgenstein (1953) put it with respect to games, "[w]e do not know the boundaries because none have been drawn…One might say that the concept 'game' is a concept with blurred edges" (68-71). They are combinatory vague because it appears to be indeterminate what *combinations of properties* are such that, were they to be fulfilled to a sufficient degree, would qualify some x for membership to an emotion kind. *Degree vagueness* emerges instead whenever it is indeterminate whether or not a property featured in the condition of membership of a kind has been fulfilled to a sufficient degree, as in the paradigmatic cases of "tall" and "bald".

Under this view, the reason why there is epistemic uncertainty in the language community as to whether, say, *calmness* and *boredom* are emotions is that "emotion" has combinatory vagueness, instantiated when we are "not able to make any sharp discriminations between those combinations of conditions which are, and those which are not, sufficient and/or necessary for application" of the category (Alston 1967, 220).

Fehr and Russell (1984) only tested the presence of disagreements among language users, and focused on only 20 target emotions 10 of which are prototypical. My view is that if we allow people to manifest their judgments of uncertainty and graded membership, "emotion" reveals itself to be a category with many borderline cases. If I take myself to be representative of the linguistic community, for example, I would have to conclude that *respect*, *startle*, *interest*, *pain*, and *lust* are examples of borderline emotions, namely emotions whose membership to the category emotion is indeterminate given our habit of language. I would be uncertain about several of them, and assign graded membership.

I am also convinced that the emotion categories subordinate to "emotion" are vague in the same sense in which "emotion" is (combinatory vague). With respect to this issue, there is some

preliminary evidence to consider. For example, Russell and Fehr (1994) considered 20 candidate subordinate categories of "anger" such as "fury", "aggravation", "exasperation", etc.. They showed that only 2 out of 28 categories subordinate to anger were judged to be instances of "anger" by 100% of the people polled, whereas there was some amount of disagreement with respect to the other 26. In the case of "torment", for example, 57% of the people polled considered it to be an instance of anger, and 43% considered it not to be an instance of anger.

Even emotions that I would personally never consider instances of anger (e.g. disgust, envy) were judged by about 70% of English speakers to be categories subordinate to anger! The picture that emerges from these data is that there is massive epistemic indeterminacy among language users concerning what things count as emotion, or as particular emotions. The data and the considerations I offered give good reasons to conclude that emotion, and its subordinates, are vague.

What I propose to retain of the epistemic approach to vagueness is a live sense of possibility that vagueness is an explanation of epistemic indeterminacy which could be wrong even when it is strongly supported. Maintaining this sense of possibility, however, is compatible with thinking that there are more evidential reasons to believe that a condition of membership that settles all cases *does not exist* for folk emotion categories, than to believe that *it exists and we do not know it*.

8.3. EMOTIONS AND HETEROGENITY

Shaver et al. (1987) presented a further study on the prototypicality of "emotion", which allows us to form a more detailed picture of the emotion hierarchy, and discuss the central topic of *heterogeneity*. They asked 100 experimental subjects to rate 213 categories that "could reasonably be considered emotion names" (1065) in terms of how good or bad an example of emotion they were. In this case, the rating system allowed for a study of prototypicality and epistemic indeterminacy at the same time. Subjects were asked to rate each category on a 4 points scale, from "I would definitely not call this an emotion" (1 point) to "I would definitely not call this an emotion" (4 points). At first blush at least, closeness to 4 indicates

prototypicality, whereas closeness to 1 indicates lack of membership. The first 98 items on Shaver et al.'s (1987) list are the following:

Average Prototypicality Ratings for "Emotion"		
1. Love (3.94.00/4.00)	**34. Lust (3.43/4.00)**	67. Gladness (3.17/4.00)
2. Anger (3.90/4.00)	**35. Disgust (3.42/4.00)**	68. Regret (3.16/4.00)
3. Hate (3.84/4.00)	**36. Hostility (3.41/4.00)**	69. Rejection (3.16/4.00)
4. Depression (3.83/4.00)	**37. Jubilation (3.41/4.00)**	70. Pride (3.14/4.00)
5. Fear (3.83/4.00)	**38. Loneliness (3.41/4.00)**	71. Gaiety (3.13/4.00)
6. Jealousy (3.81/4.00)	**39. Delight (3.40/4.00)**	72. Homesickness (3.13/4.00)
7. Happiness (3.77/4.00)	**40. Pleasure (3.40/4.00)**	73. Jolliness (3.12/4.00)
8. Passion (3.75/4.00)	**41. Tenderness (3.40/4.00)**	74. Nervousness (3.12/4.00)
9. Affection (3.72/4.00)	**42. Pity (3.39/4.00)**	75. Woe (3.12/4.00)
10. Sadness (3.68/4.00)	**43. Bitterness (3.38/4.00)**	76. Longing (3.11/4.00)
11. Grief (3.65/4.00)	**44. Disappointment (3.38/4.00)**	77. Loathing (3.10/4.00)
12. Rage (3.64.00/4.00)	**45. Humiliation (3.38/4.00)**	78. Satisfaction (3.10/4.00)
13. Aggravation (3.63/4.00)	**46. Despair (3.37/8)**	79. Hope (3.08/4.00)
14. Ecstasy (3.63/4.00)	**47. Frustration (3.37/4.00)**	80. Abhorrence (3.06/4.00)
15. Sorrow (3.62/4.00)	**48. Hurt (3.37/4.00)**	81. Insecurity (3.06/4.00)
16. Compassion (3.61/4.00)	**49. Adoration (3.36/4.00)**	82. Defeat (3.05/4.00)
17. Joy (3.62/4.00)	**50. Agony (3.35/4.00)**	83. Dread (3.05/4.00)
18. Envy (3.58/4.00)	**51. Thrill (3.34.00/4.00)**	84. Fondness (3.05/4.00)
19. Fright (3.58/4.00)	**52. Fury (3.33/4.00)**	85. Enthusiasm (3.05/4.00)
20. Terror (3.57/4.00)	**53. Remorse (3.30/4.00)**	86. Sentimentality (3.05/4.00)
21. Elation (3.55/4.00)	**54. Agitation (3.29/4.00)**	87. Hopelessness (3.04.00/4.00)
22. Guilt (3.53/4.00)	**55. Outrage (3.28/4.00)**	88. Annoyance (3.03/4.00)
23. Excitement (3.51/4.00)	**56. Resentment (3.28/4.00)**	89. Cheerfulness (3.03/4.00)
24. Anguish (3.49/4.00)	**57. Dislike (3.27/4.00)**	90. Displeasure (3.03/4.00)
25. Embarrassment (3.49/4.00)	**58. Glee (3.24.00/4.00)**	91. Melancholy (3.02/4.00)
26. Worry (3.49/4.00)	**59. Alienation (3.23/4.00)**	92. Glumness (3.01/4.00)
27. Panic (3.48/4.00)	**60. Distress (3.23/4.00)**	93. Shock (3.01/4.00)
28. Unhappiness (3.48/4.00)	**61. Enjoyment (3.23/4.00)**	94. Spite (3.01/4.00)
29. Anxiety (3.46/4.00)	**62. Relief (3.23/4.00)**	95. Suffering (3.01/4.00)
30. Desire (3.45/4.00)	**63. Gloom (3.21/4.00)**	96. Dismay (3.00/4.00)
31. Horror (3.45/4.00)	**64. Misery (3.20/4.00)**	97. Exasperation (3.00/4.00)
32. Sympathy (3.44/4.00)	**65. Euphoria (3.19/4.00)**	98. Infatuation (3.00/4.00)
33. Shame (3.4.003/4.00)	**66. Bliss (3.18/4.00)**	

Figure 9: The top 100 folk emotions. From Shaver et al. 1987

These are, at first blush at least, the categories that count as kinds of emotions in ordinary English. When asked "Would you call this an emotion?," a significant majority of the experimental subjects answer that they definitely would with respect to the 98 items reported in the table. As the decreasing numbers in parenthesis show, ordinary language users tend to consider some items (e.g. love, anger, 3.94 and 3.90 points average respectively) to be better examples of emotion than others (e.g. exasperation, infatuation, 3 points average each).

Among the items after the 98th, but still with a rating between 2.99 and 2.50, we find ire, wrath, insult, liking, neglect, astonishment, gratitude, boredom, calmness, respect, sulkiness, and indignation. Among the items in the 2.49-2.00 range we find nostalgia, modesty, vanity, exhaustion, startle. Among the items in the 1.99-1.57 range (1.57 being the lowest), we find interest, self-control, alertness, and intelligence.

What can be concluded from such data? First of all, they confirm Fehr and Russell's (1984) prototypicality results, offering a fairly similar account of the best examples of emotion. Secondly, the data suggest that there is a certain degree of consensus that the items at the top of the list are instances of emotion (e.g. love, anger, hate, depression) and the item at the bottom are not instances of emotion, or at least not definitely instances of emotion (e.g. alertness and intelligence).

But it remains unclear where the cutoff point between items that are non-prototypical emotions and items that are not emotions, or not definitely emotions, is located. For example, what does a 2.73/4.00 average rating for "awe" indicate exactly? To interpret the number, it would be useful to know the variance, namely how spread the numerical results were around 2.73. This would indicate whether 2.73 results from a significant disagreement about membership or from consensus that "awe" is a poor example of emotion. The data I discussed in section 8.2 concerning epistemic indeterminacy suggest that many of the low averages are probably to be interpreted as manifestations of epistemic indeterminacy, which I have interpreted as evidence of vagueness.

What interests me in particular about Shaver et al. (1987)'s list of emotions, however, are not the data concerning prototypicality and epistemic indeterminacy. What the list offers is an overview of the items judged by the population of speakers at large to be emotions. Figure 10 reports only emotion categories with at least a 3.00/4.00 average rating. This provision is meant to eliminate items whose status as emotions may be in question in the linguistic community. My

assumption is that, if 4 means to be definitely an emotion and 1 means to be definitely not an emotion, an average rating of at least 3 indicates the presence of a significant consensus in the population that something counts as an emotion. Not necessarily a prototypical emotion, but an emotion nevertheless. What I want to point the reader's attention to is that the list of items judged to be emotions by the linguistic community is exceptionally heterogeneous. I offer the following list of *17 dimensions of heterogeneity*, which is not meant to be complete, but only to offer a preliminary overview of the heterogeneity of the domain of phenomena people ordinarily call "emotion". All the examples I use below are taken from the list of items within the 3.00-4.00 range in Shaver et al. (1987) I reported above.

Ontological status

Items in the list vary with respect to their ontological status: some items appear to designate exclusively dispositions, others both occurrences and dispositions, and others only occurrences

For example, hostility seems to be a disposition to respond aggressively in certain circumstances, rather than an occurrence. Anger seems to comprise both dispositions and occurrences. Fright appears to designate only occurrences

Mode of onset

Items in the list vary in terms of their mode of onset: some items appear to be designate occurrence elicited quickly and automatically, and others appear to designate occurrences with a slower and less automatic onset.

For example, love and hopelessness appear to be generally elicited slowly and not automatically, whereas horror and embarrassment appear to be elicited quickly and automatically

Duration

Items in the list vary in terms of their duration: some items appear to designate short-lived occurrences, others appear to designate long-lived ones

For example, panic and disgust seems to be short-lived occurrences, whereas grief and bitterness seem to be long-lived ones

Conscious Experience

Items in the list vary in terms of whether or not they can be had without a conscious experience: some items appear to designate occurrences that can be had both consciously and unconsciously, whereas appears not liable to occur unconsciously

For example, fear and envy appear liable to being had both consciously and unconsciously, whereas shock and bliss seem to demand a conscious experience

Particular intentional objects

Items in the list vary in terms of whether or not they have particular intentional objects: some items appear to have them, and others appear not to have them.[13]

For example, depression and anxiety often do not have objects, whereas horror and jealousy generally have them

Bodily changes

Items in the list vary in terms of whether or not they tend to have a bodily underpinning of the autonomic sort: some appear to have it, and some appear to lack it

For example, anger and fear often have bodily underpinnings, whereas loneliness and regret rarely if ever appear to have them

Occurrence in infants and animals

Items in the list vary in terms of whether or not they can be had by infants and animals: some appear to be shared across species and available to infants, and others do not

For example, excitement and fright appear to be available to infants and animals, whereas guilt and envy do not

Biological primitivity

Items in the list vary in terms of whether or not they are primitive biological motivators: some appear to be, and others appear not to be

For example, lust and pleasure appear to be primitive biological motivators, whereas dread and awe do not

[13] On the other hand, I assume that emotions always have formal intentional objects, under the interpretation of formal objects I proposed in subsection 4.2.1.

Valence

Items in the list vary in terms of their valence: some appear to have positive valence, others appear to have negative valence, others appear to have mixed valence.

For example, love seems to have a positive valence, terror seems to have a negative one, melancholy seems to have a mixed valence.

Facial expressions

Items in the list vary in terms of whether or not they have facial expressions commonly associated to them

For example, panic and disgust seem to have facial expressions commonly associated to them, whereas hope and resentment do not seem to.

Prompting or inhibiting action

Items in the list vary in terms of whether they prompt or inhibit action: some appear to energize towards doing things, others appear to reduce the inclination to act

For example, disgust and jealousy seem to prompt towards action, whereas depression and gloom seem to prevent action

Priority and urgency of coping

Items in the list vary in terms of whether or not they demand urgent and prioritized coping: some appear to interrupt any other activity, others do not seem to

For example, horror and rage appear to involve immediate action which takes precedence over any pre-existing plan, whereas resentment and longing do not appear to demand immediate action

Presence of feedback in the course of execution

Items in the list vary in terms of whether or not the urgent and prioritized coping they demand can eventuate in a reflex action, or in an action which receives feedback in the course of execution: some appear to work reflex-like, and others do not

For example, fear often occurs in a reflex-like fashion, whereas frustration and alienation do not.

Attention

Items in the list vary in terms of whether or not they demand focused attention: some appear to involve redirection of selective attention on their objects, whereas others do not

For example, fury and ecstasy appear to demand attention to be focused on their objects for the entire duration of the episode, whereas melancholy and remorse do not

Integration with long term planning

Items in the list vary in terms of whether or not they are integrated with long term planning: some appear to commonly bring about impulsive actions, whereas others appear to be more integrated with long term planning

For example, panic appears to bring about impulsive actions, whereas resentment seems to be integrated with long term planning.

Complexity of appraisal

Items in the list vary in terms of the cognitive complexity of the appraisal which brings them about: some are elicited by an appraisal process which presupposes and involves higher cognitive capacities, others appear to be elicited by a low-level appraisal process

For example, guilt, remorse, and seem to require the ability to engage in complex cognitive processing, whereas panic, lust, olfactory disgust do not

Machiavellian elicitation

Items in the list vary in terms of whether or not their elicitation mechanism has a prominent Machiavellian dimension: some are elicited by an appraisal which manifests sensitivity to the payoff expected from engaging in them, whereas others do not

For example, anger and jealousy are often elicited in order to affect the behavior of an interactant, whereas insecurity and shock appear less geared towards influencing interactants advantageously

These are 17 dimensions of heterogeneity affecting the category "emotion". I collected them to show how multi-dimensional the conceptual hyperspace occupied by the category "emotion" actually is. The reader has certainly noticed that the seventeen dimensions of heterogeneity are dimensions around which some of the main attempts to characterize emotions in the last 2,500 years have been centered. I have discussed most of them in the previous seven chapters of this dissertation. In other words, the seventeen dimensions are the output of many centuries of "looking and seeing", as Wittgenstein (1953) would put it, what emotions as ordinarily understood have in common. None of the dimensions I considered, as it turns out, individuates a feature that is necessary for being an emotion as ordinarily understood.

At this point, we have two options. One is to keep trying, and the other is to give up on the idea that the category of "emotion" is individuated by a set of individually necessary and jointly sufficient conditions. Given how long the history of unsuccessful attempts to find them, the appropriate inductive inference seems to me that individually necessary and jointly sufficient conditions for something to count as a folk emotion have not been recovered *because there are no such conditions*. Famously, this is what Wittgenstein concluded with respect to games, even though his inductive basis was much more limited than the one I rely on, as it only comprised his own attempt to find out what all games have in common.

Wittgenstein's well-known conclusion was the following:

> Consider for example the proceedings we call "games". I mean board-games, card-games, ball-games, Olympic games, and so on. What is common to them all?-Don't say: "The must be something common, or they would not be called games"- but look and see whether there is anything common to all.-For if you look at them you will not see something that is common to all, but similarities, relationships, and a whole series of them at that…we see a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail. I can think of no better expression to characterize these similarities than "family resemblances"…[a]nd I shall say: 'games' form a family (Wittgenstein 1953, § 65-67)

Paraphrasing Wittgenstein, I will say that also "emotions" form a family. Notice that this is an aspect of the analogy between emotions and games different from the one I discussed in section 8.2. There, I reported that Wittgenstein was convinced that there are borderline cases of

game, namely cases such that it is indeterminate whether or not they definitely are or are not games. I concluded that the same holds for the category emotion" (as well as for categories subordinate to it such as "anger" and "fear"), and presented some preliminary empirical evidence on epistemic indeterminacy to back that up (Wittgenstein relied on introspection to make the vagueness point about games). What I am saying now is something different, namely that the items which are definitely emotions form of family, characterized by a "complicated network of similarities overlapping and criss-crossing".

I am convinced that the same is true of categories subordinate to "emotion". For example, there seem to be instances of *fear* which are occurrences, and instances which are dispositions, instances which are elicited quickly, and instances which are elicited slowly, instances which involve conscious experiences, and instances which are unconscious, instances which involve bodily changes, and instances which lack them, instances which are short-lived and accompanied by facial expressions, and instances which are long-lived and not accompanied by facial expressions, instances which have great urgency, and instances which do not, instances which work as reflexes, and instances which comprise feedback mechanisms in the course of execution, instances which are elicited by simply appraisal mechanisms, and instances which are elicited by complex ones, and so on. As I will put it, also "fear", "anger", "guilt", "shame', "disgust" each form a "family", whose members, once again, share a complicated network of similarities overlapping and criss-crossing.

Under the view I am proposing, folk emotion categories designate families of items which share a family resemblance, and have borderline members. However, a student of emotions cannot afford stopping here, and say about emotions only that they share a complicated network of similarities overlapping and crisscrossing. The question is: What options are available for theorizing about emotions?

8.4. CONCLUSION

I surveyed the available empirical literature on emotion concepts, to try and understand what sorts of items are ordinarily considered to be emotions, or particular emotions. I reported that competent English speakers deem some items to be better and some items to be worse examples of emotions, which should not surprise us given how widespread prototypicality phenomena are with respect to all kinds of lexical categories. What is more interesting is that there is evidence of widespread epistemic indeterminacy when it comes to folk emotion categories. Competent English speakers appear to be uncertain or disagree about whether or not some items count as instances of emotion, or of a particular emotion.

I also surveyed the domain of items English speakers consider to be definite examples of emotion, and detected the presence of a massive heterogeneity within such domain. I reported seventeen dimension of heterogeneity, gleaned from the long history of failed attempts to define the emotions. The best interpretation of the empirical data I presented, I concluded, is that folk emotion categories lack a condition of membership comprising individually necessary and jointly sufficient conditions of membership, and have blurred edges.

9. WHAT ARE THE DESIDERATA FOR A THEORY OF EMOTIONS?

The emotion theorists we have surveyed up to now tend to ask "What is an emotion?" or “What is anger?” without a clear understanding of what counts as getting the answer right. An account of what counts as getting the answer right must begin from the conclusion I reached at the end of the last chapter, namely that folk emotion categories are vague and characterized by family resemblance. I want to argue that there are two different projects in which an emotion theorist may be engaged with respect to such categories. The first is what we may call the Folk Emotion Project, which aims to offer a *descriptive account* of the conditions of membership of folk emotion categories such as “emotion” and “anger”. The second option is to engage in what I call the Explicating Emotion Project. This project aims to offer *explicative accounts* (or explications) of folk emotion categories which aim to transform them into explicative categories endowed with fruitfulness relative to a certain set of theoretical objectives. Uncertainty about the desiderata of these two projects, and ambiguity about which one is being pursued, strike me as the two biggest methodological obstacles to progress in the history of emotion theory. The aim of this chapter is to remove such obstacles, and get clear on the ground rules of the activity of answering questions of the form “What is an emotion/anger?”.

9.1. THE FOLK EMOTION PROJECT

9.1.1. Folk emotion categories as cluster categories in a fuzzy hierarchy

One of the projects in which a theorist answering questions such as “What is an emotion?” or “What is anger?” may be engaged is the Folk Emotion Project (FEP). Its chief purpose is to

offer what I call a *descriptive account* of folk emotion categories, namely an account of their condition of membership which is compatible with the empirical facts I have illustrated in sections 8.1, 8.2 and 8.3. In a nutshell, the folk emotion theorist must shed light on the family resemblance condition of membership of a vague category. Such condition must explain why there is more than one way to qualify as an emotion/anger - another way to say that emotions form a family - and why there are borderline cases of emotion/anger.

Notice that this is a substantive intellectual task, which is certainly not fulfilled by dictionary definitions. The Webster Dictionary, just to pick an example, tells us that an emotion "is a state of feeling" or that anger is "a strong feeling of displeasure and usually of antagonism". But these descriptive accounts are *definitions*, namely sets of individually necessary and jointly sufficient conditions, and we have already concluded that no account with such logical form can capture the condition of membership for folk emotion categories. The two specific definitions offered by the Webster dictionary comprise conditions which are neither individually necessary nor jointly sufficient. They are not jointly sufficient because many things other than emotions are states of feeling (e.g. pain, nausea) and many things other than anger are strong feelings of displeasure and usually of antagonism (e.g. jealousy). They are not individually necessary because many things which are instances of emotion do not meet the proposed definition (e.g. unconscious emotions) and many things which are instances of anger do not meet the proposed definition (e.g. self-righteous anger can be mild and somewhat pleasant).

The question is: How should folk emotion categories descriptively accounted for? What counts as a good descriptive account for a vague category whose instances only share a complicated network of similarities overlapping and criss-crossing? My view is that the appropriate logical form for it is that of a *cluster account*. I propose the following as a general description of what I shall call a *cluster condition of membership*:

> A category has a cluster condition of membership when membership to the category is a matter of fulfilling enough properties from a cluster set $\{P_1,..,P_n, Pj_?\}$

I call *cluster account* a sentential account of a condition of this sort, and *cluster category* a category with a cluster condition of membership. A cluster condition of membership states that

being a category member is fulfilling *enough* of the properties from the cluster set $\{P_1,..,P_n, Pj_?\}$. The fundamental features of the cluster condition I described are two.

Firstly, there is more than one way of fulfilling it, which is what accounts for Wittgenstein's "family" idea. Often many different combinations of properties from the cluster set will be jointly sufficient for membership, namely such as to meet the threshold for "enough". This will create a "complex network of similarities overlapping and criss-crossing" among category members.

Secondly, *indeterminacy* is built into the very condition of membership for the category, which is what accounts for Wittgenstein's "blurred edges" idea. There are two sides to this indeterminacy. On the one hand, the key term "enough" admits of borderline cases. Secondly, the identity of the cluster set $\{P_1,..,P_n, Pj_?\}$ is not fully determinate, in the sense that there may be some property Pj such that it is not determinate whether or not it belongs to the cluster set. Belonging to the cluster set is in effect being a component of at least one combination of properties which are "enough" for membership (without any proper subset of such set being enough for it), and there will be indeterminacy as to whether or not some properties ever achieve that distinction. For the sake of simplicity, I will disregard this aspect of indeterminacy in what follows, and work under the assumption that there is no indeterminacy at least as to the identity of $\{P_1,..,P_n\}$, which may be the case for some emotion categories.

Notice that what a cluster condition of membership states is not that with respect to borderline cases we are ignorant about whether or not enough properties have been fulfilled, but rather that there is no fact of the matter concerning whether or not enough properties have been fulfilled. No amount of investigation will settle the issue, because – under the understanding of vagueness I am presupposing - our habits of language have left it an open question how the issue ought to be settled.

The admissibility of borderline cases, however, should not be construed as global indeterminacy. There will be cases such that the cluster condition of membership has definitely been fulfilled and cases in which it has definitely been violated. For example, fulfilling *all* properties in the cluster set of any cluster category is definitely having fulfilled enough properties, and fulfilling none of them is definitely not having fulfilled enough properties. This means, incidentally, that the cluster properties must be jointly sufficient for membership, and that fulfilling none of them must be sufficient for non-membership.

The *domain of indeterminacy* of cluster category, on the other hand, will differ from category to category. Some categories will be such that *any* proper subset of the cluster set is a borderline case: whenever not all properties from the cluster set have been fulfilled, it is neither definitely the case that the condition of membership has been fulfilled nor definitely the case that it has been violated. In such case, the cluster set will be the only sufficient condition of membership for the kind. Other categories will be such that there will be several proper subsets of the cluster set that are jointly sufficient for membership, so that the domain of indeterminacy will be more restricted. These differences may be hinted at by choosing other terms in lieu of "enough", such as "most", "many", "some", "sufficiently many", and so on.

Cluster categories differ also in another important respect, namely whether or not the cluster set contains individually necessary properties. For example, it may well be the case that being voluntary is a necessary condition for being a game, but that being a game is being voluntary and fulfilling enough properties from the cluster set (which includes the necessary property). In this respect, Wittgenstein (1953) was wrong in thinking that what makes games a "family" is that there is no properties all games must share in order to be games. They could be a family even if such properties existed, as long as there are several ways of fulfilling the cluster condition of membership for the category.

My central hypothesis is that folk emotion categories have a cluster condition of membership, in the sense that membership to them is a matter of fulfilling enough properties from a cluster set $\{P_1,..,P_n\}$.[14] Call this the *cluster hypothesis* for folk emotion categories. The cluster hypothesis (CH) explains very well the empirical data I collected in sections 8.1, 8.2 and 8.3.

It explains why definitions of emotions in terms of individually necessary and jointly sufficient conditions have so far eluded emotion theorists, who have tried to define the emotions for centuries without success. According to CH, this is because the only thing which is necessary and sufficient when it comes to folk emotion categories is fulfilling enough properties from the cluster set. It also explains why there is a great deal of heterogeneity among instances of folk emotion categories, in the sense that there are many ways to fulfill enough properties from a

[14] As I pointed out before, the cluster set itself may have borderline members, but I won't pursue this point in what follows

cluster set. Finally, CH explains why there are borderline cases of emotions, as sometimes it is simply indeterminate whether or not enough properties from the cluster set have been fulfilled.

If I am right, the primary tasks for a *folk emotion theorist* is to shed light on the cluster set associated with the folk category "emotion" (and its subordinates), find out what properties in the cluster set are individually necessary (if any), and find out what proper subsets of the cluster set are sufficient conditions of membership (if any).

This is a substantive task, which is to be carried out by investigating empirically emotion concepts along the lines pioneered by the psychologists I have discussed in chapter 8. We certainly know what the candidates for the cluster set are, namely what in the introduction I called the *marks of emotionality*, namely the prototypical components involved in instances of prototypical emotions such as anger, fear, disgust and so on. Such components comprise an appraisal, a suite of physiological responses, a conscious experience, and a behavioral action tendency manifested by physical actions, mental actions and expressions. But it is unclear whether these really are the properties of the cluster set as I defined it, whether any of them is necessary for something to qualify as a folk emotion, and what subsets of them are sufficient for something to qualify as a folk emotion.

What the folk emotion theorist is ultimately after is to shed light on what I call the Fuzzy Hierarchy of Folk Emotion Categories. It looks something like this:

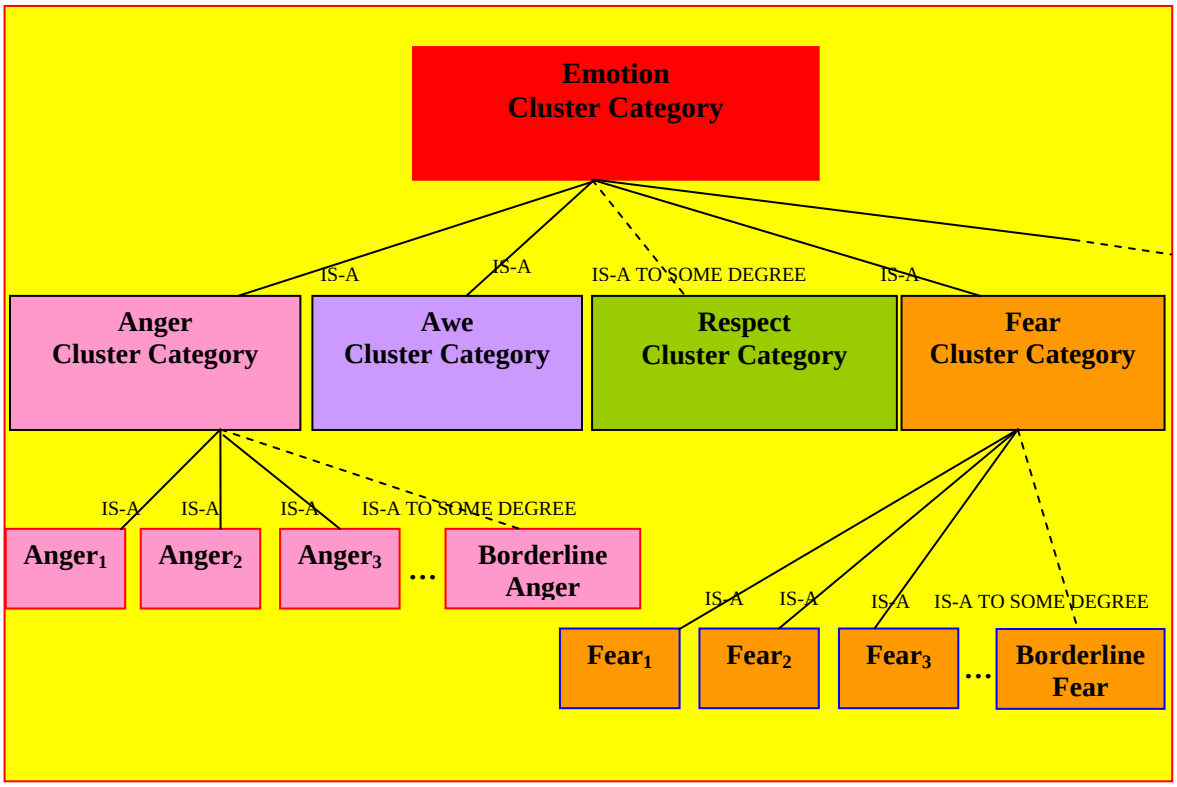


Figure 10: The fuzzy hierarchy of folk emotion categories

In the hierarchy described above, *anger*, *awe* and *fear* are *classically subsumed* under *emotion*, because every instance of *anger*, *fear* and *awe* is an instance of *emotion*, but there are instances of *emotion* which are not instances of *anger*, *fear*, or *awe*. On the other hand, *respect* is *fuzzily subsumed* under *emotion*, because instances of *respect* are borderline instances of *emotion*, but at least one instance of emotion (e.g. an anger instance) is definitely not an instance of respect. We need the distinction between these two types of subsumption if we want to allow for cluster folk emotion categories. Such categories will enter relations of subsumption with other folk emotion categories, but they will not be relations of subsumption as classically understood, namely relations of set inclusion.

In figure 11, I indicated classical subsumption with a full line labeled "__IS-A__", and fuzzy subsumption with a dotted line labeled "__IS-A__TO SOME DEGREE". Each of the emotion cluster categories will be characterized by a cluster set which the folk emotion theorist must discover. There will be *many* emotion kinds which are *classically subsumed* under *emotion*, namely many categories all the instances of which fulfill enough properties of the cluster set for "emotion". Some of them will be *prototypical emotion categories*. Their instances are not only

definitely instances of emotion, but also very good examples of *emotion*. On the basis of the evidence described in section 8.1, I suggest that the group includes, among others, *anger, fear, happiness, sadness, love, hate, guilt,* and *sadness*.

Other kinds classically subsumed under emotion will be *non-prototypical emotion categories*, namely kinds whose instances are definitely instances of emotion, but not prototypical ones. This group may include, among others, *awe, gloom, spite, infatuation*. There will also be emotion kinds which are *fuzzily subsumed* under *emotion*, in the sense that with respect to their instances it is neither definitely the case that they are emotions nor definitely the case that they are not. The group of *borderline emotion categories* includes, among others, *respect, boredom, calmness, startle, interest* and many others. The instances of such categories are such that it is indeterminate whether or not they fulfill enough properties from the cluster.

Each of the prototypical, non-prototypical and borderline folk emotion categories will in turn subsume other categories. For example, there will be several categories of anger - $anger_1$, $anger_2$, $anger_3$, etc. - which are classically subsumed under anger, in the sense that their instances fulfill enough properties from the cluster set of anger to qualify as anger.

Species of anger are generated by the specific way in which the cluster condition of membership is met and by the specific way the cluster properties are fulfilled. Some of these kinds will be *prototypical anger categories*, and others will be non-*prototypical anger categories*. Prototypical anger categories tend to have most or all of the cluster properties, and tend to fulfill them in a prototypical way. For example, a prototypical anger category would includes instances of anger which involve appraisal of slight, autonomic changes such as increased heart beat, blood pressure and trembling, an attack tendency manifested through expressions such as fixed stare, eyes widened, and bared teeth, physical behaviors such as screaming and hitting, mental behaviors such as focusing attention on the object of anger and plotting further harm in the future, and a negatively valenced feeling.

Some of the categories subsumed under anger have a name in the English language, but there certainly are more categories of anger that there are names for them. An example of named anger category could be *rage*, which tends to have all of the characteristics I listed above. On the other hand, there will be several instances which fulfill enough properties from the cluster set for anger to count as borderline anger, but not enough to count as anger.

In other words, it will be indeterminate whether or not they are anger, and no amount of investigation will settle the issue. I include all such borderline instances in the *borderline anger kind*, which is fuzzily subsumed under *anger*. The same is going to be true of *fear*, which will classically subsume *prototypical* and *non-prototypical* fear kinds, and fuzzily subsume a *borderline fear kind.*

9.1.2. Are folk emotion categories natural kinds?

I am not interested in the Folk Emotion Project, even though I acknowledge that it is an interesting and intellectually coherent project. This means that I will not try to offer a characterization of the cluster set which individuates "emotion", or any particular emotion such as "anger", nor try to reconstruct the precise shape taken by the fuzzy hierarchy of folk emotion categories.

The reason is that, whatever the cluster condition of membership for folk emotion categories may be, we already know that the items which fulfill it display the sort of massive heterogeneity and vagueness I documented in sections 8.2 and 8.3. But categories of this kind are not the sorts of categories suitable for scientific investigation. To develop scientific theories of emotions, which is what I am interested in, we need some degree of precision, and most importantly we need to individuate a domain of phenomena which share a dimension of scientifically interesting similarity.

Another way to formulate this point is to say that folk emoting categories are not natural kinds. This thesis has been made popular by Griffiths (1997), but has recently been attacked from a variety of fronts (e.g. Nussbaum 2001, Charland 2002, Prinz 2004c; see Griffiths 2004b for replies). Griffiths (2004b) uses the term natural kind "to denote categories that admit reliable extrapolation from samples of the category to the whole category", and argues that "[i]deally, a natural kind should allow very reliable predictions in a large domain of properties" (235). This approach is broadly inspired by Boyd's (1991) analysis of natural kinds as *homeostatic property clusters kinds*.

Boyd was interested in natural kinds such that some natural mechanism underlies the satisfaction of the condition of membership on the part of kind members. The condition of membership presupposed by Boyd is similar to what I called a cluster condition of membership, namely a cluster of properties such that being a kind member amounts to fulfilling *enough* of

them. This being the case, there is no fixed set of necessary and sufficient properties that all natural kind members must fulfill. What characterizes Boyd's kinds is that the satisfaction of enough of the cluster properties is the result of a *causal* process:

> There are a number of scientifically important kinds (properties, relations, etc.) whose natural definitions are very much like the property-cluster definitions postulated by ordinary language philosophers except that the unity of the properties in the defining cluster in manly causal rather than conceptual. The natural definition of one of these *homeostatic property clusters kinds* is determined by the members of a cluster of often co-occurring properties and by the ("homeostatic") mechanisms that bring about their co-occurrence (Boyd 1991, 141, emphasis in original).

What makes homeostatic property cluster kinds *natural* is that it is, as it were, up to nature what properties "cluster" in virtue of causal mechanisms. Whereas we are free to define kinds by the combination of any cluster of properties we wish, only some clusters will have a causal basis. For example, we are free to generate kinds by establishing sets of properties that, say, stars in the sky must fulfill. We can define a kind as being formed by stars in a region of space that, seen from the earth, looks like a lion. In such case, most likely no causal mechanism will underlie the satisfaction of the condition of membership for the kind.

On the other hand, when we define the kind "spiral galaxy", whose condition of membership is that of "exhibiting a central nucleus or barred structure from which extend concentrations of matter forming curved arms" (Merriam-Webster's Collegiate Dictionary), being member of the kind has a causal basis, in the sense that a natural common cause must have brought the stars belonging to the spiral galaxy to gather in that particular configuration we call spiral galaxy.

Boyd points out that the attempt to define kinds whose condition of membership is fulfilled in virtue of a causal mechanism is significant only when we are engaged in inductive and explanatory projects:

> In defining a kind we should be required to defer to the world just in case and to the extent that reference to the kind in question is to be part of an inductive or explanatory project. In cases in which our concerns are largely with the establishment of workable

> conventions for non-inductive practice, deference to the world should be largely unnecessary (1991, 140)

The reason is that when we want to engage the world of nature in our explanatory and inductive projects, we must defer to the causal structure of such world. As Boyd puts it, "successful induction and explanation always require that we accommodate our categories to the causal structure of the world" (139). Since the causal structure of the world is something we discover as we go along with induction and explanation, natural kinds will receive their ultimate definitions "a posteriori in deference to nature rather than nominally" (139).

Under this broad characterization of natural kinds, Griffiths (2004b) argues that folk emotion categories are not natural kinds. His central point is that "it is unlikely that all the psychological states and processes that fall under the vernacular category of emotion are sufficiently similar to one another to allow a unified scientific psychology of emotion" (2004c, 233). Griffiths extends this thesis to specific folk emotion categories, suggesting that vernacular categories such as anger and love are unlikely to be natural kinds. Notice that this is not to say that emotions have nothing in common, or that what they have in common cannot be discovered. It is simply to say that instances of folk emotion categories do not share the sorts of properties "that are the focus of investigation in psychology and the neurosciences".

Griffiths (1997, 2004b) argues that the only types of emotions for which we have evidence of naturalness are affect programs (or basic emotions), namely biologically based and pan-cultural suites of short-term, coordinated and automated responses which includes measurable physiological changes, stereotyped facial expressions and action tendencies (see chapter 5). Ekman's (1999b) most recent list of such programs comprises surprise, amusement, anger, contempt, joy, disgust, embarrassment, excitement, fear, guilt, pride in achievement, relief, sadness/distress, satisfaction, sensory pleasure, and shame. As I pointed out in 5.1.3, however, so far we only have evidence for basicness with respect to anger, fear, disgust, sadness, surprise and joy.

Differently from Ekman, Griffiths acknowledges that affect programs are not the only things that qualify as folk emotions, but he points out that they are sufficiently different from affect programs to require a distinct scientific psychology. Griffiths (2004b) suggests two main research paths along which we may develop theories suitable for other natural kinds of emotions. Some emotions appear to require "responding in a more cognitively complex way to more highly

analyzed information" (2004b, 236) than it is the case for basic emotions. This seems to be the case for emotions such as envy, jealousy, moral indignation, resentment and others. Griffiths proposes to call such emotions *higher cognitive* or *complex emotions*.

On the other hand, on some occasions emotions appear to "involve an internalized cultural model of appropriate behavior". Griffiths suggests that this appears to be the case for emotions such as "going postal", which seem to follow a script "derived from real or fictional incidents that are culturally salient" (2004b, 236) (see my discussion of this case in subsection 6.3.2). Griffiths proposes to call such emotions "socially sustained pretenses".

Griffiths' primary concern is not to develop detailed accounts of either higher cognitive emotions or socially sustained pretenses, but only to point out that these emotions cannot be understood as affect programs or combinations of affect programs. This being the case, we cannot extrapolate to the folk category of emotion the scientific discoveries made about affect programs. Similarly, since there are instances of folk categories such as anger and disgust which are not affect programs, we cannot extrapolate to the folk categories of anger or disgust the scientific discoveries made about affect program anger and affect program disgust. In a nutshell, neither the folk category of emotion nor the folk categories of specific emotions such as anger, disgust, shame, etc. are natural kinds.

Griffiths considers a possible objection to his argument, namely that all instances of emotion which are not instances of affect programs may be reduced to blends of affect programs. Griffiths (1997) has a number of responses to this challenge. Firstly, he believes that the elicitors of some emotions such as jealousy and moral indignation cannot be reduced to combinations of elicitors of affect programs. Secondly, he points out that many emotions are not short-lived, whereas affect programs are. Thirdly, he states that many emotions "do not have immediate behavioral and physiological consequences" (1997, 102) as affect programs do. Fourthly, he argues that some emotions are "highly integrated with complex, often conscious cognitive processes", whereas affect programs appear to have many of the properties of modules (e.g. informational encapsulation).

Griffiths' conclusion is that many instances of folk emotion cannot be reduced to combinations of affect programs. Consequently, folk emotion categories are neither natural kinds nor reducible to combinations of basic emotions which instead are natural kinds.

As I pointed out before, Griffiths' thesis has been criticized by several emotion theorists in recent times. The standard criticism is that Griffiths failed to appreciate that there are things *all* emotions share by virtue of which they are natural kinds. I will consider only Prinz's (2004c) version of the criticism, because it offers a vivid portrayal of a general problem.

Prinz's case for holding that Griffiths' thesis is wrong is summarized in the following passage:

> Each [emotion] is structurally analogous. Each is simply a perception of a patterned bodily change. Even emotions that we acquire by blending [between basic emotions] have this simple structure. They are simply perceptions of blended bodily patterns. Some emotions are attained by adding conceptually sophisticated judgments to out elicitation files, but this does not alter their structure. Elicitation files are content-determining causes of our emotions, not constituent parts. And all emotions have elicitation files that can contain judgments, as well as perceptual representations. Thus, hybrid theories are wrong. All named emotions are very much alike. All have the same internal structure, and all bear the marks of both nature and nurture (Prinz 2004c, 85-86)

I already criticized Prinz's Neo-Jamesian theory because of its unwarranted assumption that all emotions are perceptions of bodily changes (see section 7.2). I will further criticize it in chapter 10 because of the theory of intentionality it presupposes, and argue that it fails to account for the crucial motivational dimension of emotions. Here, I want to argue that Prinz's (2004c) argument for the naturalness of the folk category "emotion" would not work even bracketing all such worries. Let us assume for the sake of argument that "[a]ll emotions are embodied appraisals under the causal control of calibration files", where "calibration file" is another name for what I earlier described as appraisal. Consider fear. According to Prinz, all instances of fear are perceptions, conscious or unconscious, of the bodily changes of fear broadly understood (see subsection 7.1.1). Such perception was set up to be set off by items contained in the fear calibration file, such as loud noises and loss of support.

However, fear can be calibrated through higher cognitive processes as well. For example, fear can be generated by the judgment that a meteorite will hit the earth in two years. The fear calibration file, in other words, contains items resulting from calibration of fear elicitors, which

reflect different levels of cognitive complexity and cultural differences. But what about higher cognitive emotions such as guilt or shame? Prinz's response is that "[h]igher cognitive emotions are either blends of two basic emotions (just as martinis are blends of two spirits), or combinations of basic emotions and cognitive elaborations (just as screwdrivers combine a spirit and a fruit juice)". Prinz concludes that emotions "share a common essence" by virtue of which they are natural kinds, namely being *all* embodied appraisals.

The problem is that this argument does not address the real issue, which is that members of folk emotion categories are extremely heterogeneous. Let us assume for the sake of argument that, by sufficiently stretching the notion of perception of bodily changes, and allowing for calibration files to range from cognitive primitive to cognitively complex forms of information processing, we manage to conclude that "[a]ll emotions are embodied appraisals under the causal control of calibration files", either in the form of basic emotions or in the form of combinations of basic emotions. It would still be the case that such embodied appraisals manifest the seventeen dimensions of heterogeneity I described in section 8.3.

To summarize, some embodied appraisals would be occurrences and others would be dispositions, some would be elicited quickly, and some would be elicited slowly, some would involve conscious experiences, and some would be unconscious, some would involve autonomic changes, and some would not, some would be short-lived and accompanied by facial expressions, and some would be long-lived and not accompanied by facial expressions, some would have particular intentional objects, and some would lack them, some would occur in infants and animals, and some only in adult humans, some would prompt action, and some would inhibit it, some would work as reflexes, and some would involve action tendencies, and so on. What needs to be demonstrated is that there is a dimension of scientifically interesting similarity which holds across all such differences, and this is not something we can find in Prinz's theory.

What Prinz ultimately tells us is that all emotions have the same internal structure, namely being embodied appraisals, and that all embodied appraisals bear the marks of both nature and nurture. He adds to such qualifications that "[a]ll emotions seem to involve overlapping brain structures, and all can be affected by the same clinical conditions" (2004c, 77). Let us assume all such things are true of emotions. The fact is that being embodied appraisals, bearing the marks of both nature and nurture, involving overlapping brain structures and being affected by the same clinical conditions does not individuate a kind such that it is likely that "all the psychological

states and processes that fall under [it] are sufficiently similar to one another to allow a unified scientific psychology" (Griffiths 1997, 233).

Prinz's (2004c) argument simply does not address the problem raised by Griffiths, offering nothing more than a characterization – problematic in its own right – of what all instances of emotion have in common by virtue of which we call them emotions in the vernacular sense. But showing that the somatic theory "subsume[s] anything that deserves to be called an emotion (Prinz 2004a, 49) does not amount to showing that anything that deserves to be called an emotion forms of natural kind.

I am convinced that Griffiths' (1997, 2004b) argument is here to stay. I would not formulate it, however, in terms of the distinction between affect programs such as anger and fear, higher cognitive emotions such as guilt and envy, and socially sustained pretenses such as "going postal". This formulation carries two risks. The first is to suggest that higher cognitive emotions and socially sustained pretenses form natural kinds of emotions, as affect programs do. The second is to suggest that the each folk emotion category falls neatly into one of these three kinds of emotion.

Griffiths does not hold any of these two views. He mostly describes higher cognitive emotions and socially sustained pretenses in terms of the fact that they are *not* affect programs, without explicitly arguing that they "are sufficiently similar to one another to allow a unified scientific psychology". Moreover, he is open to the possibility that the same kind of emotion – say anger - may have instances that count as affect programs, higher cognitive emotions and socially sustained pretenses.

Griffith's interpreters, however, often take answering the claim that emotions are not natural kinds to amount to showing that there is unity in the emotion domain despite the tripartition between affect programs, higher cognitive emotions and socially sustained pretenses. We have seen the impact of this assumption on Prinz's (2004c) argumentative strategy.

Holding that folk emotion categories are not natural kinds does not require holding that there are any specific natural kinds of emotions other than basic emotions (e.g. higher cognitive emotions and social pretenses). For that matter, it does not even require holding that basic emotions are natural kinds. All we need to say is that it is very unlikely that the "complex network of similarities overlapping and criss-crossing" which characterizes folk emotion

categories is a network of similarities allowing for the formulation of scientific explanations and predictions that hold for all instances of the folk category.

9.2. THE EXPLICATING EMOTION PROJECT

The problem with folk emotion categories, I argued, is that our habits of language qualify a very heterogeneous set of items as definite instances of them, and do not settle with respect to several items whether or not they are in fact instances of the category. If the objective of a theorist is to individuate with precision an interesting dimension of similarity within the domain of emotions, it is not advisable to work with folk emotion categories. Notice that this would be true even if the Folk Emotion Project had already been successfully completed. Having a cluster account for folk emotion categories would make explicit what makes something a folk emotion (or a particular folk emotion), but would not eliminate the dimensions of heterogeneity which characterize it (I detected seventeen such dimensions), nor the vagueness of the category. In fact, eliminating vagueness would qualify as a shortcoming for a descriptive account, whose objective is to mirror, rather than reduce or eliminate, the vagueness of the folk category whose condition of membership is being described.

What a theorist needs to do in these circumstances is to transform folk emotion categories so as to make them suitable to his or her theoretical purposes. My own objective is to transform folk emotion categories so as to make them suitable for the purposes of scientific psychology. But I want to begin by offering a general discussion of the desiderata which govern what I call the *Explicating Emotion Project*. Its chief purpose is to offer an *explication* of folk emotion categories, which is roughly speaking an account of what is a "good thing to mean" by them relative to certain theoretical purposes. The job of an emotion theorist engaged in explication is not to shed light on the family resemblance condition of membership of vague folk emotion categories, as is the case for the Folk Emotion Project. Rather, the purpose is to create a new category which is useful to certain purposes, but similar enough to the old category to count as explicating it. I will try to shed light on explication by introducing and developing Carnap's pioneering account.

9.2.1. Carnap's account of explication developed

According to Carnap's (1950) broad characterization, "[t]he task of explication consists in transforming a given more or less inexact concept into an exact one or, rather, in replacing the first by the second. We call the given concept (or the term used for it) the *explicandum*, and the exact concept proposed to take the place of the first (or the term proposed for it) the *explicatum*" (3). Let us call *explicans* the account of the condition of membership for the *explicatum* or *explicative kind*. The question we have to answer is: What desiderata should a good explication fulfill? The general intuition is that a good explication preserves a significant portion of the meaning of the explicandum, and makes up for whatever is lost by conferring upon the explicatum epistemic virtues lacked by the explicandum.

Carnap's account of the desiderata of explication reflects the theoretical purposes for which the notion was introduced, namely scientific purposes. We will have to transform the Carnapian account slightly to endow it with more generality, because scientists are not the only ones who engage in explication.

Here is a summary of what Carnap takes the desiderata of explication to be (Carnap used the term "concept" as I am using the terms "category" or "kind"):

> If a concept is given as explicandum, the task consists in finding another concept as its explicatum which fulfils the following requirements to a sufficient degree.
>
> 1. The explicatum is to be similar to the explicandum in such a way that, in most cases in which the explicandum has so far been used, the explicatum can be used; however, close similarity is not required, and considerable differences are permitted.
>
> 2. The characterization of the explicatum, that is, the rules of its use (for instance, in the form of a definition), is to be given in an exact form, so as to introduce the explicatum into a well-connected system of scientific concepts.
>
> 3. The explicatum is to be a fruitful concept, that is, useful for the formulation of many universal statements (empirical laws in

> the case of a nonlogical concept, logical theorems in the case of a logical concept).
>
> 4. The explicatum should be as simple as possible; this means as simple as the more important requirements (1), (2), and (3) permits (Carnap 1950, 7)

Let us label the four desiderata as *similarity*, *exactness*, *fruitfulness*, and *simplicity*. The desideratum of *similarity* states that the explicatum must be usable "in most cases" in which the explicandum has been used, but "considerable differences are permitted" and "close similarity is not required". Expressions such as "most cases", "considerable differences", and "close similarity" are vague expressions, which admit borderline cases. In other words, there will be situations in which it is not determinate whether or not the differences between the way in which explicandum and the explicatum are used are similar enough to achieve explication.

What Carnap (1950) leaves implicit are the *uses* on which the similarity between explicandum and explicatum is to be grounded. As I interpret him, what he is referring to is the way in which the *terms* designating explicandum and explicatum are used in (the same) language. Under this view, whether or not similarity is achieved is contingent upon the degree to which the explicatum term is interchangeable with the explicandum term in such a way that the sentence obtained after substitution maintains the properties it had before substitution. More precisely, I think Carnap is referring to the preservation of the *truth value of declarative sentences*. Under this view, explicandum term and explicatum term are similar to the extent that they are *interchangeable salva veritate* in a favored set of declarative sentential contexts. There must be many such contexts in which interchangeability is preserved, although "close similarity" is not required and "considerable differences are permitted". [15]

This is the sense in which an explicatum does not aim for *alikeness in meaning* with the explicandum, which is what a definition as generally understood aims to bring about. An explication is not an attempt to capture *all and only* the meaning of the explicandum, but only to capture a good portion of such meaning – the one embodied by the degree of interchangeability achieved – while fulfilling other desiderata at the same time. This is the sense in which we can

[15] For simplicity of reference, I will drop the qualifier "term", but keep assuming the interpretation of similarity in use I just presented.

speak of an explication as stating a "good thing to mean" by the explicandum, rather than what the explicandum means.

Carnap (1950) argues that the explicatum must be characterized in "exact form", so that it can be introduced "into a well-connected system of scientific concepts". He also argues that the explicatum must be fruitful, namely "useful for the formulation of many universal statements". Finally, Carnap (1950) states that the explicatum must be "as simple as possible". I will disregard the desideratum of simplicity in what follows, because it may very well be that an explicatum "as simple as the more important requirements (1), (2), and (3) permit[]" is in fact extremely complex. The epistemic virtue of being as simple as it is required to fulfill requirements other than simplicity, namely similarity, exactness and fruitfulness, sounds fairly trivial and is certainly not central to determine the quality of an explication.

Carnap's account of the desiderata of explication points us in the right direction, but needs some modifications. What is good about it is that it emphasizes what I take to be the two main objectives of an explication, namely offering a characterization of the explicatum in terms of precise rules of use and offering a characterization of the explicatum which endows it with fruitfulness. However, I think these desiderata need not be pursued at the same time, and they do not have to be understood exclusively in terms of scientific exactness and scientific fruitfulness.

On the contrary, Carnap (1950) argues that an explication is good just in case it fulfills *both* exactness and fruitfulness, along with similarity, "to a sufficient degree". Moreover, he assumes that the rules of use of the explicatum must have the sort of exactness which introduces the "explicatum into a well-connected system of scientific concepts" and that what makes the explicatum fruitful is that it is useful "for the formulation of many universal statements (empirical laws in the case of a nonlogical concept, logical theorems in the case of a logical concept)".

My view is that there are perfectly good explications which only aim for and achieve exactness understood as precision in the rules of use, and perfectly good explanations which only aim for and achieve fruitfulness understood as usefulness with respect to an open range of purposes, not necessarily scientific ones. For example, a government may want to issue laws regulating the free exercise of "religion", and may wish to exclude borderline cases of religion with respect to which it is indeterminate whether or not the law applies. This would be a fine

reason to explicate “religion”, but there would be no other aim than exactness for the purposes of law.

On the other hand, a philosopher may want to achieve both exactness and fruitfulness, but for the purposes of philosophy. Philosophical accounts of categories such as “knowledge”, “perception”, and “inference” very rarely “introduce the explicatum into a well-connected system of scientific concepts”, or formulate “universal statements” about the explicatum. The point is, they do not aim to, and there is no reason why they should. Given that fruitfulness is a discipline-relative notion, I think we ought to assess explications *relative* to the theoretical purposes of a particular discipline, not simpliciter.

It is even questionable that scientists themselves have as strict an understanding of exactness and fruitfulness as Carnap (1950) does. For example, there is debate as to whether any universal statements hold true the special sciences. Asking an explication to generate an explicatum embeddable in “many universal statements” appears to be asking too much even for scientific purposes. I will reformulate the idea of fruitfulness by equating it with usefulness in the formulation of generalizations (explanatory, predictive, etc.) which meet the standard of adequacy of the relevant discipline, without making any further assumptions on the nature of such generalizations.

In summary, my reformulation of Carnap’s (1950) desiderata for explication takes the following shape:

> If a category C is given as explicandum relative to the purposes of discipline D, the task consists in finding another category C* as its explicatum which fulfils the following two requirements to a sufficient degree:
>
> 1. The explicatum C* is to be similar to the explicandum C in such a way that, in most cases in which C has so far been used, C* can be used; however, close similarity is not required, and considerable differences are permitted.
>
> 2. The characterization of the explicatum C*, that is, the rules of its use is to be given in an exact form <u>and/or</u> the explicatum C* is to be a fruitful concept, that is, useful for the formulation of the explanatory and predictive generalizations characteristic of D

One of the big risks when introducing explications is that of generating ambiguity, namely multiplicity of meaning. For example, if we were to call "knowledge" an explicatum obtained from the explicandum "knowledge", the term "knowledge" would designate both the explicandum and explicatum. This would be a problem, because whereas the explicandum is the set of things which fulfill the cluster accounts individuated by folk knowledge theorists, the explicandum is the set of things which fulfill the proposed explicans. A superscript (knowledge*), a subscript ($knowledge_1$), or a new name which wears the explicative relation on its sleeves (knowledge-how) would be equally good ways to avoid the problem of ambiguity.

9.2.2. Explicating emotions

One of the projects in which a theorist answering questions such as "What is an emotion?" or "What is anger?" may be engaged is the Explicating Emotion Project (FEP). Its chief purpose is to transform folk emotion categories such as "emotion" and "anger" into *explicative kinds* which are similar in use to the folk categories, and decrease their vagueness and/or increase their fruitfulness. The explicating emotion theorist starts from the same folk categories from which the folk emotion theorist starts. Differently from a descriptive account, however, an explication does not aim to capture all and only the meaning of folk emotion categories.

Consider "emotion", the folk category I propose to explicate in the next chapter by means of what I call the Urgency Management System (UMS) theory of emotions. To explicate it successfully would be to construct a notion of "emotion*" which is similar to "emotion" in such a way that in most cases in which "emotion" has so far been used, "emotion*" can be used. Given the gloss I put on the idea of *use*, what this means it that "emotion*" has to be interchangeable salva veritate with "emotion" in most linguistic contexts. But, crucially, "emotion*" can be a good explicatum for "emotion" even if there are *many* linguistic contexts in which "emotion*" is not interchangeable with "emotion". This is because "considerable differences in use" between "emotion*" and "emotion" are permitted by the very ground rules of explication. Emotion theorists who were to criticize an explication of emotions because it encounters some ordinary language Type 1 and Type 2 counterexamples (see section 7.1) would simply not have understood what an explication is.

Now, the payoff(s) which "emotion*" must bring to the table in order to be a good explication are an increase in precision and/or fruitfulness with respect to "emotion". But, as I

argued in subsection 9.2.1, we can't assess an explication of emotion in the abstract, but we must do so relative to the theoretical objectives of a specific discipline that determines a reference class for what is fruitful. The explicative project I am most interested in is that of characterizing explicative emotion kinds that are useful for the purposes of scientific psychology.

Before getting started, I want to emphasize two important features of explication I would like the reader to keep in mind as she or he goes through the theory I propose in the next chapter.

The first is a general problem in establishing whether or not an explication is a good one. Carnap pointed out that since the explicandum "is not given in exact terms [and] since the datum is inexact, the problem [of explication] itself is not stated in exact terms". This being the case, "if a solution for a problem of explication is proposed, we cannot decide in an exact way whether it is right or wrong". The main point I take Carnap to be making is that will often be hard to tell whether or not explicatum and explicandum are sufficiently similar to one another to instantiate the explication relation.

Carnap (1950) criticized philosophers for trying to provide explications without having understood how the explicatum is used in ordinary language. Carnap scolds for example those philosophers who ask "'What is causality?', 'What is life?', 'What is mind?', 'What is justice?', [and] immediately start to look for an answer without first examining the tacit assumption that the terms of the question are at least practically clear enough to serve as a basis for an investigation, for an analysis or explication".

This is the spirit in which I have offered an extended discussion of the empirical literature about emotion concepts, and argued that it is the appropriate starting point for an explication of emotions. If we do not know how folk emotion categories are used in ordinary language, we won't be able to aim for similarity in use with respect to them, and one of the two key requirements for a good explanation will go unfulfilled.

What I will try to formulate in the next chapter is an explication of folk emotion categories which fulfills the requirement of similarity in a clear way, namely one which stays clear of the worry raised by Carnap concerning borderline cases of similarity.

At the same time, the account I propose does not aim to avoid all ordinary language counterexamples, but only to avoid enough of such counterexamples for it to be clear that the differences in use between "emotion" and "emotion*", although "considerable", are of the permitted" kind.

The other important feature of explication I want to emphasize is that it is intrinsically *pluralistic*. Given any folk emotion category C, there will be innumerable explicata C*, C**, C*** which achieve similarity in use with C in a favored set of linguistic contexts, and fulfill exactness and/or fruitfulness relative to a given set of theoretical purposes. In other words, to understand the desiderata of explication is to understand that there are many "good things to mean" by the same explicandum, which are not synonymous with one another nor with the explicandum, but manage to decrease vagueness and/or increase fruitfulness in their own, distinctive ways.

Consequently, what I propose is not the only conceivable theory of emotions, but what I take to be a good theory of emotions relative to the purposes of scientific psychology.

I conclude with a figure which summarizes the two projects an emotion theorist may be engaged in:

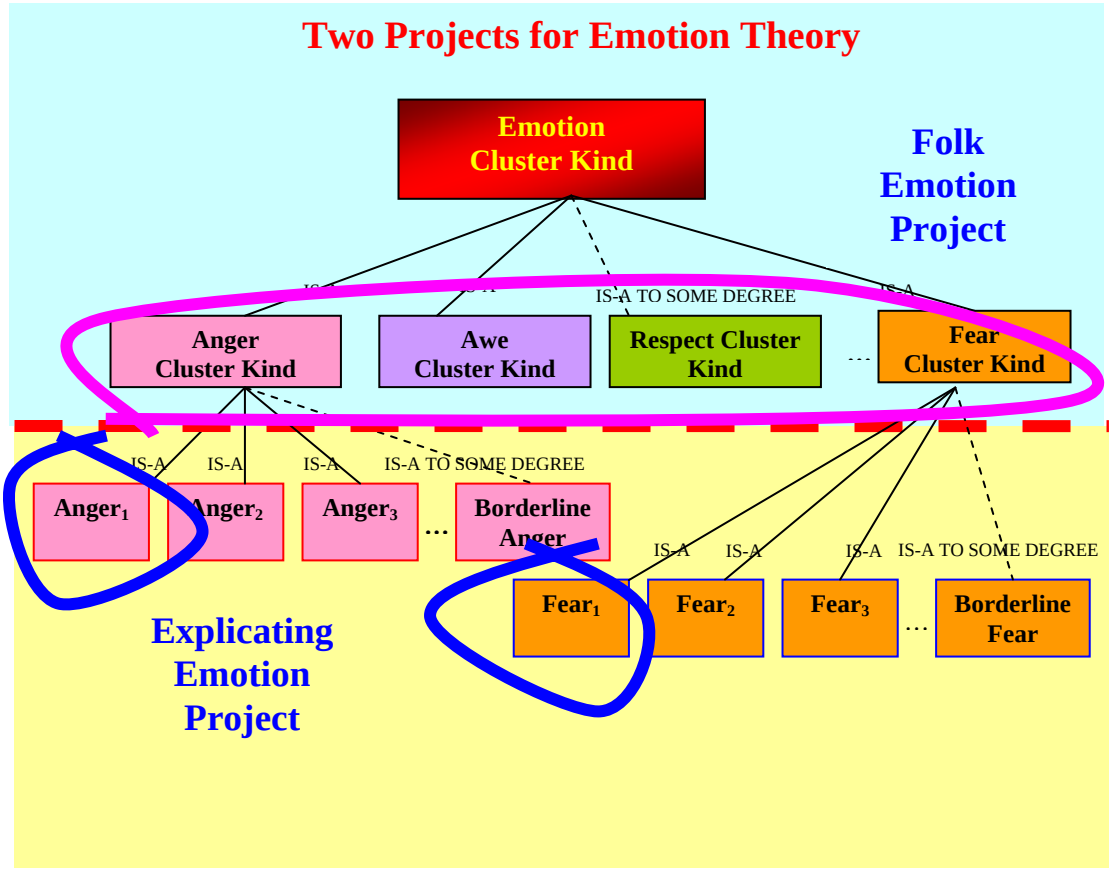


Figure 11: Two projects for emotion theory

Whereas the Folk Emotion Project aims to formulate cluster accounts of folk emotion categories which account for all of their instances, the Explicating Emotion Project aims to formulate explications for such categories which do not apply to the entire vernacular domain. For example, a cluster account of "emotion" would have to characterize a condition of membership which accommodates the facts that anger, awe and fear are emotions, and respect is a borderline emotion. An explication of "emotion" has not such requirement. It can capture a dimension of similarity shared by certain kinds of angers ($anger_1$) and certain kinds of fears ($fear_1$), but not shared by some other kinds of anger ($anger_2$) and fear ($fear_2$), and not shared by any kind of awe and respect. Even under such circumstances, it may be the case that the explication is a good one, as long as it preserves similarity in use with the explicandum, and fulfills the desiderata of reducing vagueness and/or increasing fruitfulness.

9.3. CONCLUSION

The history of emotion theory is a long sequence of attempts to individuate a subset of marks of emotionality such that anything that deserves to be called an emotion fulfills them. I pointed out that this notion of deservingness has generally been ambiguous between two interpretations, namely that of *theoretical fruitfulness* and that of *ordinary language compatibility*. Emotion theorists have been unclear about what sort of deservingness they were pursuing, in part because of the widespread assumption that what emotion terms "mean" coincides with what is a "good thing to mean" by them relative to the purposes of a theory.

I have argued that emotion theorists should decisively divorce the two projects which, as I have argued, have been run together for most of the history of emotion theory. One is the project of capturing what can rightfully be called an emotion in ordinary language (Folk Emotion Project), and the other is the project of capturing what is worth calling an emotion relative to the purposes of a given theory (Explicating Emotion Project).

In the next chapter, I will offer a new theory of emotions in the context of the Explicating Emotion Project.

10. EMOTIONS AS URGENCY MANAGEMENT SYSTEMS

In this final chapter I aim to offer a new theory of emotions. My objective is to explicate the ordinary notion of emotion so as to individuate a precisely characterized and potentially fruitful theoretical construct I call "umotion". [16] The neologism "umotion" is meant to signal that what I am offering is not a descriptive account of "emotion", but rather an account to be assessed in light of the desiderata for explication I discussed in section 9.2. It is also meant to remind the reader that the fundamental feature of umotions is *U*rgency. More precisely, I characterize an "umotion" as an "urgency management system", and I label the theory I construct around such systems as the "urgency management system" (UMS) theory of emotions.

What is the take home message of the UMS theory? In a nutshell, it is that an umotion is a special type of superordinate system which activates and manages an urgent action tendency by coordinating the operation of a cluster of cognitive, perceptual and motoric subsystems. Crucially, such a superordinate system has a proper function by virtue of which it acquires a special kind of intentionality I call pragmatic.

The theoretical construct of umotion emerges from the integration of two core ideas. The first idea is that umotions are special *action control structures* devoted to the management of situations which involve the pursuit of high priority goals. I borrow this idea from Nico Frijda's (1986)'s theory of emotions as action tendencies, which I develop in several directions. The second idea is that umotions are *intentional pushmi-pullyu representations*: they acquire normativity from having proper functions, and they combine descriptive and directive purposes

[16] The term "Umotion" was suggested to me by Paul Griffiths

into an undifferentiated whole. These ideas are adapted from Ruth Millikan's (2004) theory of intentionality, but applied to umotions in novel ways.

The UMS theory of emotions, therefore, provides a new account of the *vehicles* of emotional representation, understood as urgency management systems, and a new account of the *representation relation*, understood in teleosemantic terms. I argue that the theory I provide is better equipped than either cognitivism or Neo-Jamesianism, currently the two most popular theories of emotions, to fruitfully explicate what the emotions are relative to the purposes of scientific psychology. The differences between the UMS theory and such theories are both methodological and substantive. Differently from cognitivists and Neo-Jamesians, I do not claim that my theory captures everything that deserves to be called an emotion in ordinary language. I am explicitly in the business of explicating emotion, and all I argue for is that there is "similarity in use" between the folk category "emotion" and the explicative kind "umotion" I introduce. Also, I do not claim that my theory captures the only interesting explication of emotion we can come up with. In fact, I am convinced that there are many interesting explicative projects in emotion theory other than the one articulated by the UMS theory.

Substantively, my theory starts from the assumption that the *evaluations* embodied by umotions and the way umotions *feel* are phenomena to be understood in light of the impact umotions have on action. Cognitivists and Neo-Jamesians consider instead evaluations and feelings to be what emotions are essentially, leaving the crucial relation between emotion and action in the background. Under the view I propose, evaluations and feelings are important but not essential components of the urgency management systems with which emotions are identified.

The argumentative strategy I employ comprises three steps. The first is to illustrate Frijda's (1986) theory of emotions as action tendencies, which is the main inspiration for my own theory. The second step is to offer a detailed account of the features of umotions as urgency management systems. The third step is to explain in what sense urgency management systems have pragmatic intentionality. I will conclude by explaining why I take the UMS theory to be a good theory of emotions.

10.1. DEVELOPING FRIJDA

The idea I start from is a simple one, namely that when organisms emote they are *inclined to act* in ways which are related to the type of emotion they are having in a non-arbitrary, although non-deterministic, fashion. For example, a person who is angry with Alexandra and a person who is in love with her are *inclined* to act towards her in very different ways, although what they specifically end up doing may on occasion be similar (e.g. trying to find her). The idea that emotions involve *impulses for action* is of course an old one. Aristotle, for example, stated that anger "may be defined as an impulse, accompanied by pain, to a conspicuous revenge for a conspicuous slight directed without justification towards what concerns oneself or towards what concerns one's friends" (*Rhetoric* 2. 2 1378a31-1378b1).[17] The question is: what sort of impulses do emotions generate?

The best answer to this question I know of is offered by Frijda (1986). The central thesis of Frijda's theory of emotions goes as follows:

> Emotions…can be defined as modes of relational action readiness, either in the form of tendencies to establish, maintain, or disrupt a relationship with the environment or in the form of mode of relational readiness as such (Frijda 1986, 71)

The central notion here is that of a mode of *relational action readiness*, which can exist either in the form of an *action tendency* or in the form of *readiness as such*. Frijda borrowed the notion of a *tendency* from Magda Arnold's (1960) theory of emotions (see section 4.3). According to Arnold (1960, 182), an emotion is the "felt tendency toward anything intuitively appraised as good (beneficial), or away from anything intuitively appraised as bad (harmful)". To appraise something intuitively is for Arnold (1960) to appraise it in a way which is *immediate* and *direct*, roughly in the sense that it does not involve slow and laborious thought processes. As Arnold notes, an elephant appraising whether or not a certain ground will sustain its weight, or a ball player appraising whether or not a flying ball can be caught will make such

[17] As I argued in chapter 2, anger is the only emotion with respect to which Aristotle makes explicit the impulse component

appraisals "intuitively".[18] But what exactly counts as a *tendency*? Neither Arnold nor Frijda offer us an explicit account of tendencies, so I will provide a working one.

I understand *tendencies* as *dispositions* in the philosophical sense, namely properties associated with a typical *manifestation* and a typical set of *triggering circumstances*. Although the precise logical relation between dispositions and subjunctive conditionals is controversial (see Mumford 1998 for discussion), I will assume that dispositions are associated with a subjunctive conditional of the form "if triggering circumstances T were fulfilled, then the bearer would display manifestation M with probability p".[19] It is important for my subsequent analysis that it is not assumed that a disposition is necessarily manifested when the triggering circumstances for it are fulfilled: some dispositions are probabilistic, namely such that their associated manifestation follows their triggering circumstances with a probability of less than 1.

Ryle (1949) remarked that an expression such as "tends" suggests that "it is a good bet that it will be…the case" (131). This idea can be expressed in terms of subjunctive conditionals by saying that a tendency is a disposition such that the manifestation follows the triggering circumstances with a significant probability (there will clearly be borderline cases). Let us work with this rough and ready account, and ask whether emotions are the sort of felt tendencies described by Arnold (1960). It is quite clear that not all tendencies towards things or away from them are emotions, whether or not they are felt. For example, I may intuitively appraise, say, a banana as good and feel a tendency towards eating it, or intuitively appraise a business proposal as bad and feel a tendency towards rejecting it, without an emotion being instantiated in either case. If we want to identify emotions with tendencies, we need to qualify what kinds of tendencies we are talking about, because many tendencies are clearly not emotions.

As a first approximation, Frijda's (1986) theory is that emotions are *action tendencies*,[20] a notion he characterized as follows:

---

[18] I propose a caveat to the idea that emotional appraisal is always immediate and direct in sub-section 10.2.3

[19] I follow Elizabeth Prior (1985) in thinking that the ascription of dispositions generally presupposes a set of *background circumstances* C in which the disposition is assumed to hold. For example, a given object X counts as having the dispositions of fragility, solubility, or inflammability not simpliciter, but given a set of background circumstances C. The same object may be, say, fragile given a background temperature of -170°C, but not fragile given a background temperature of 30°C.

[20] They are action tendencies when they are not states of *readiness as such*, the other species of the "relational action readiness" genus contemplated by Frijda

> Action tendencies are states of readiness to execute a given kind of action. A "given kind of action," and thus an action tendency, is defined by its end result aimed at or achieved. That end result can be inferred from behavior; the basis of inference is the behavior's flexibility…Action tendency is readiness for different actions having the same intent. One action tendency is readiness for attacking, spitting, insulting, turning one's back, or slandering, whichever of these appears possible or appropriate at a given moment; a different action tendency is readiness to approach and embrace, fondle, look at avidly; or say sweet things, again according to what the circumstances favor (70-71)

This passage states that having an action tendency amounts to being in a particular state of readiness, namely a readiness "to execute a given kind of action". What *kind* of action it is will depend on what *kind of* "end result" is being pursued. In the passage quoted at the beginning of this section, Frijda spoke of action tendencies as "tendencies to establish, maintain, or disrupt a relationship with the environment". I will call *relational goal* (or *purpose*) the goal associated with an action tendency, namely the goal towards which the tendency is assumed to be flexibly directed. Frijda takes flexibility to be the primary manifestation of goal-directedness, relying on the assumption that, as Woodfield (1976, 51) once put it, "sensitivity to changed conditions is a good sign of goal-seeking ability in general". The relational goal of the action tendency, Frijda suggests, is to be inferred from "behavior's flexibility", namely from the fact that an action tendency can be manifested "according to what the circumstances favor" or to what "appears possible or appropriate at a given moment". What makes *different actions* manifestations of the *same action tendency* is their "having the same intent", namely aiming to fulfill the same relational goal. The relational goal is not to be understood as whatever end-state is finally reached once the action tendency is manifested, but in terms of an end-state which is *supposed* to be reached when a given action tendency is manifested. The way in which I will account for this essential normative aspect of teleological processes is in terms of Millikan's notion of a proper function, which grounds the possibility of genuine goal-failure in terms of a history of selection (see below).

Before discussing proper functions, there is a major problem with Frijda's theory we need to address. The problem is that an emotion cannot possibly be *any* state of readiness to choose "different actions having the same intent". For example, someone may be currently ready to

choose actions fulfilling a given relational goal *if certain further circumstances were fulfilled*, but not be in an emotional state. For example, I am now in a state of readiness to attack, insult, slander, etc. Howard Stern *if I am threatened by him*. This does not mean that I am currently in a state of anger towards Howard Stern, although I am disposed to get into one if provoked. The sort of readiness Frijda presupposes must be readiness *in the current circumstances*.

A further problem is that there are forms of current readiness to achieve a certain relational goal which have nothing to do with emotions. For example, when the commuter train I have been waiting for finally arrives, I am ready to hop on it in the current circumstances to fulfill the goal of going to work, but there need not be anything emotional about my hopping. This problem is addressed by Frijda (1986) through the introduction of the important idea of *control precedence*:

> Action tendencies have the character of urges or impulses. - Action tendencies - and action readiness changes generally - clamor for attention and for execution. They lie in waiting for signs that they can or may be executed; they, and their execution, tend to persist in the face of interruptions; they tend to interrupt other ongoing programs and actions; and they tend to preempt the information-processing facilities…Evidently, then, action tendencies are programs that have a place of precedence in the control of action and of information processing. We therefore say: Action tendencies - action readiness changes generally - have the feature of *control precedence* (78)

> Emotion, as action readiness state or as emotional action, has action control precedence in two senses. It can interrupt other processes and block access to action control for other stimuli and other goals; it invigorates action for which it reserves control and invests that control with the property of indistrability or persistence (460)

These quotes reveal that Frijda is ambiguous about the nature of action tendencies. He had previously described them as mere states of readiness, but he now characterizes them as states of readiness *endowed with control precedence* (or *urgency*). Since not all action tendencies "have the character of urges", I propose we distinguish between action tendencies with and without control precedence. The former are what Frijda ultimately takes emotions to be, and they are the

sorts of action tendencies I will focus on from now on when I speak of *urgent action tendencies*. Frijda's key insight is that action tendencies with control precedence are those which "clamor for attention and for execution". This "clamoring" strikes me as being both the primary, and the least explored, of the marks of the emotional. We must now ask: what exactly does it mean for an action tendency to clamor for attention and execution? In the passages I just quoted, Frijda introduces a number of distinct aspects to this "clamoring".

The first is that states of readiness with control precedence "tend to interrupt other ongoing programs and actions". This is an idea Herbert Simon (1967) put at the center of his theory of emotions as *interrupt systems*. Simon's basic point was that there is a "close connection between the operation of the interrupt system and much of what is usually called emotional behavior" (35), and that "an interruption mechanism, that is, emotion, allows the [information] processor to respond to urgent needs in real time" (38).[21] Although interruption characterizes the operation of many action tendencies with control precedence, it does not characterize the operation of them all. On some occasions, "urgent needs" are in fact connected to the pursuit of an "old" goal. For example, I may be engaged in trying to get a store clerk to give me a refund for a purchase and get angry in the course of my interaction with her. In such case, I would still be engaged in the attempt to get a refund, but I would do so in an angry way, without interruption of other ongoing programs and actions. Wisely, Frijda only claims that urgent action tendencies "tend" to interrupt.

The second element introduced by Frijda (1986) is that states of readiness with control precedence "lie in waiting for signs that they can or may be executed". I interpret this passage as an attempt to suggest that an emoter is not only ready to fulfill a certain relational goal in the sense of being prepared for it, but in the stronger sense of trying to find *means* to it. This aspect seems to me a central aspect of what gives emotions their urgency, and it needs to be further developed. What is missing from Frijda's formulation is a proper characterization of the fact that, far from "lying in

[21] The main problem with Simon's (1967) theory is that he failed to provide an account of what happens *after* an emotion "interrupts", other than saying that "the response program may, and often will, activate the autonomic response system" (35) and that there may be the generation of "subjective feelings" "produced, in turn, by internal stimuli resulting from the arousal of the autonomic system" (35). The activation of the autonomic systems and its attendant feelings, however, are accessory as far as Simon's theory is concerned. But to say that emotions are interrupt systems which lead to the pursuit of urgent needs in real time is just to give the headline for a theory of emotions. Providing the actual theory demands explaining how exactly the pursuit of urgent needs is supposed to take place.

waiting", emotional action tendencies generate an *active exploration* of the environment aimed at finding ways to achieve or maintain a certain kind of relationship with it.

A third component is that the search for actions that fulfill the relational goal is accompanied by bodily and mental *preparations* for such execution. As Frijda puts it, an emotional action tendency "invigorates action for which it reserves control" and it "tend[s] to preempt the information-processing facilities" and "block[s] access to action control for other stimuli and other goals". The invigoration of action takes place through the priming of the body for action, whereas "preemption" of information processing facilities and "blocked access to action control" suggest that emotions involve the coordinated operation of the entire cognitive architecture, which becomes geared towards a given relational goal.

A fourth element is that, once the execution process is under way, it has some degree of "indistrability", in the sense that the "execution…tend[s] to persist in the face of interruptions". I interpret this point as indicating that there is some degree of inertia in an urgent action tendency once it begins manifesting itself (but I will disregard this particular aspect of control precedence in what follows).

The picture that emerges from these passages is that emotions are *action control structures* which prioritize the pursuit of certain relational goals, prepare for their fulfillment both mentally and physically, and protect the execution of actions aimed at fulfilling them from possible interferences. I take this to be a very promising vantage point for theorizing about emotions, and I will construct my own theory of emotions around it.

Frijda (1986) discusses two importantly different ways in which emotions can acquire control precedence:

> The system constituting emotion is constructed, it appears, so as to allow for control precedence: There would seem to be two ways to account for this feature. The first: There exist certain action programs that in reflex-like fashion are linked to the mismatch (and potential match) signals under concern; these links might provide for built-in control precedence…The second: match and mismatch signals are responsible. The signals involved are highly persistent; they are loud and claim attention; and the system recognizes their claim upon control and their nature as calling for change or for continuing along present lines (92)

The first form of control precedence characterizes *reflexes*, which generate the automatic and immediate pursuit of a relational goal upon detection of a certain stimulus. The second form of control precedence characterizes instead *action tendencies proper*, which clamor for attention and execution along the lines I just described. Frijda realizes that "[t]o the extent that action programs are fixed and rigid, the concept of action tendency loses much of its meaning…action readiness only exists to the extent that inhibition can block action execution" (83). This passage suggests that reflexes are tendencies in an inverted commas sense at best. Most importantly, action tendencies have, and reflexes lack, a feed-back mechanism which can guide execution in real time, and eventually lead to inhibition if there are no opportunities for successful execution. This being said, I follow Frijda (1986) in thinking that we should consider reflexes to be limiting cases of action tendencies, in which the manifestation automatically and unfailingly follows the triggering circumstances without feedback in the course of execution. When I speak of an urgent action tendency, I will generally refer to an action tendency proper, but I allow the notion to comprise reflexes as a special case. The primary focus of my theory, I emphasize it, is on urgent action tendencies which are *not* reflexes.

Frijda's central thesis is that a great many emotions can be identified with action tendencies with control precedence, under the broad understanding of such notions I articulated. For example, Frijda suggests that the "action tendency of anger is interpreted as a tendency to regain control or freedom of action – generally to remove obstruction" (88). Fear is characterized as the action tendency of "avoidance", associated with the relational goal of achieving one's "own inaccessibility". Disgust is the action tendency of "rejecting", characterized by the relational goal of "removal of object". Surprise is the action tendency of "interrupting", which the relational goal of "reorientation". All of these action tendencies can be flexibly manifested, depending on the circumstances. For example, the action tendency of anger - call it *attacking* - can be manifested by spitting, insulting, turning one's back, slandering, and so on, as the circumstances of its elicitation allow and recommend.

Although the construct of an urgent action tendency is central to Frijda's account, he is careful to add that "[n]ot all emotions are action tendencies" (71). For example, sadness does not appear to be an urgent tendency to act, but if anything the opposite, namely the absence of impulses to do much of anything. On the other hand, joy often does not appear to be an urgent

tendency to do anything in particular, but rather a general eagerness to engage in many different activities. How can we account for such cases of emotion?

Frijda (1986) does so by introducing two further theoretical constructs, namely that of a "null state" and that of an "activation mode". These are both manifestations of "modes of relational readiness as such", the other species of the genus "modes of relational action readiness" with which Frijda identifies emotions. According to Frijda, sadness is a *relational null state*, namely a state of "explicit absence of relational activity" (22). Joy, on the other hand, is the "manifestation of *free activation*" (38, emphasis added). With these caveats, Frijda believes that the cases of sadness and joy fall within the purview of his theory, in the sense that "null states, activation modes, and action tendencies proper, all are modifications of action tendency in a general sense: they all represent modes of readiness, unreadiness included, for relational action" (71).

There is an air of ad hocness to Frijda's account of sadness and joy. My view is that the similarity between sadness and joy on the one hand, and anger, fear, disgust or surprise on the other does not lie merely in the fact that these are all changes in readiness to act. The main similarity lies in the fact that they all have *control precedence*, under a suitably broad understanding of such notion. At the heart of the notion of "control precedence", I argue, is the idea that emotions exert prioritized control on what the organism will do next. Generally, this control is geared towards the pursuit of a specific goal to be achieved with priority. The cases of sadness and joy are different, in the sense that they generally occur when a certain goal is either no longer achievable (sadness) or has already been achieved (joy).

But there is a sense in which sadness still exerts a strong influence on what the organism will do next, in the sense that it curtails the ability to pursue goals in general. Joy, on the other hand, has its own way of clamoring for attention and execution, often characterized by the absence of a specific goal pursuit but geared towards an open set of possible goals.

I will come back to the cases of sadness and joy later on, but I want to emphasize right away that the theory I am proposing does not specifically aim to accommodate the cases of sadness and joy. It is built around different exemplars of emotions, namely emotions which prioritize the pursuit of specific goals when they are still achievable and prior to achieving them. Sadness and joy are (at best) special cases with respect to the theory I have to offer. As I will argue later on, failing to accommodate some of the uses of the explicandum is a common

feature of good explications, and it is not *as such* a reason to reject a theory offered in the context of the Explicating Emotions Project.

Now, is Frijda's theory convincing, once we complement it with the caveats I introduced so far? As I see it, the theory still has four major limitations. The first is that it fails to clarify in much detail the nature of the clamoring for attention and execution that characterizes urgent action tendencies, especially with respect to the *active search* for means to the relational goal. The second limitation is that it restricts the phenomenon of emotion to the emergence of an action tendency with control precedence, leaving in the background the way the action tendency is managed through time by means of a feedback mechanism. The third limitation is that although urgent action tendencies occur very often in the context of a social transaction, their communicative dimension is barely mentioned in Frijda's theory. The fourth and most important limitation is that Frijda's (1986) theory is silent on the issue of intentionality. When cognitivism emerged in the 1960s from the ashes of the behaviorist theory of emotions, early cognitivists argued that thinking of emotions as behavioral predispositions (or feelings) failed to account for their intentionality. Any theory aiming to assimilate emotions to particular kinds of action tendencies must be able to explain what sort of intentionality they have. My task in the next few sections is to get rid of these four major limitations, and in the process construct a satisfactory theory of emotions as urgency management systems.

10.2. EMOTIONS AS URGENCY MANAGEMENT SYSTEMS

10.2.1. My account in a nutshell

Emotions are, in slogan form, "urgency management systems" (UMS), in short "umotions". This identity claim is offered as an explication: I argue that "umotion" is a good explicatum with respect to the explicandum "emotion" relative to the purposes of scientific psychology. Four central ideas inspire the UMS theory of emotions. The first, borrowed from Frijda's (1986) work, is that the mark of the emotional is *urgency*, namely priority in the control of action. I will understand action in a broad sense, which includes physical actions, expressions, and mental actions. Unlike Frijda, I claim that the management of an urgent action tendency, rather than

merely its activation, is part of umotion. For example, I take the physical movements of an organism experiencing fear not to be something other than fear, but rather one of the components of fear itself, in the same sense in which the physiological discharges of fear may be.

The second idea is that the urgency characteristic of umotions is due to their being *superordinate systems* with access to practically all cognitive, perceptual and motoric subsystems available to the organism. The notion of a "superordinate program" plays a prominent role also in the theory of emotions proposed by evolutionary psychologists Tooby and Cosmides (1990, 2000). Their theory, however, lacks a clear account of the way in which the resources of the organism are organized in emotion, and it assumes that "programs" are evolved modules, a position I reject (see 5.2.1).

The principle of organization I propose is that of the *management* of an urgent action tendency. As I understand it, the notion of management is broad: it includes modes of management focused on the urgent action tendency itself (call it *tendency-focused* management) and on the relational goal of the tendency (call it *goal-focused* management). For example, tendency-focused forms of management may comprise the attempt to *re-appraise* the event that led to the urgent action tendency, to make sure that it really has the features that make the urgent action tendency an appropriate response to it. It may also comprise the attempt to regulate the physiological discharges associated with the urgent action tendency, practicing controlled breathing, exercising control on muscle tension, and so on. Although tendency-focused management is an important phenomenon, I leave it in the background in my analysis, and concentrate on goal-focused management. [22]

I distinguish three broad functional components in the (goal-focused) management of a state of urgent action readiness, which I call *preparation*, *action*, and *communication*. *Preparation* has

[22] The distinction between tendency-focused management and goal-focused management is inspired by Lazarus' (1991, 2001) distinction between *emotion-focused* and *problem-focused coping*, but it differs from it in several respects. Lazarus (2001, 45) characterizes *coping* as "the effort to manage psychological stress", whereas I speak of *management* as the effort to manage *any* urgent action tendency, whether or not it involves stress. Also, Lazarus claimed that whereas problem-focused coping involves "acting to change the person-environment relationship", emotion-focused coping "involve[s] mainly thinking rather than acting" (112). I make no such assumption. As I understand it, goal-focused management often eventuates in mental actions (e.g. thinking about killing someone in anger rather than actually killing him), and tendency-focused management in physical actions (e.g. working on one's breathing patterns to try to calm down in fear). The key issue is whether efforts are directed at regulating the process by which the relational goal is pursued (goal-focused management) or the tendency itself (tendency-focused management). The distinction is not cut-and-dry, since many efforts appear to be focused at the same time on the goal and on the tendency (e.g. re-appraising the eliciting event to both establish whether the tendency is appropriate to it and how best the relational goal can be fulfilled).

to do with getting ready to execute one of the actions that share the relational goal of the action tendency, *action* has to do with executing a particular action, and *communication* has to do with broadcasting emotional signals throughout *preparation* and *action*.

The third central idea of my account is that no specific set of subsystems is activated in all cases of umotion, in the sense that distinct urgent action tendencies, and the same action tendency at different times, may recruit different groups of organismic subsystems. I allow for the possibility that an emotion may be instantiated with many different forms of *preparation*, *action*, and *communication*, each involving various possible combinations of subsystems. What is non-negotiable is only that an urgent action tendency is active and is being managed. Depending on the tendency at hand and its specific eliciting circumstances, this task can be carried out in a variety of ways.

The fourth idea is that the intentionality of umotions cannot be understood by invoking the sorts of representations commonly used in the philosophy of emotions, namely merely descriptive representations (e.g. beliefs) and merely imperative representations (e.g. desires). This is because umotions do not divorce the aim of telling emoters what is the case, as descriptive representations do, from the aim of directing action, as imperative representations do. Rather, umotions collapse descriptive and directive functions into an undifferentiated whole. I will try to capture this idea by using Millikan's (1996, 2004) theory of *pushmi-pullyu representations*. Since this will require an extended discussion, I first characterize urgency management systems and then offer an account of their intentionality. I emphasize that umotions as I understand them are *urgency management systems endowed with intentionality*. What they represent, I will argue, is an essential part of what they are.

10.2.2. Umotion defined

Here is the central thesis of the Urgency Management System (UMS) theory of emotions:

> An umotion E is a superordinate system, generally activated by an appraisal, which:
>
> > (a) controls a cluster of organismic subsystems whose synchronized operation instantiates, and manages through time, an action tendency T_E with control precedence,

(b) has a pragmatic object, which describes the conditions of pushmi-pullyu appropriateness PP_E for E

Under this view, what type-identifies an umotion E is a combination of an urgent action tendency T_E and of a set of conditions of pushmi-pullyu appropriateness PP_E for it. In turn, such conditions of appropriateness are related to the proper function of the mechanism activating the urgent action tendency. Roughly speaking, this proper function will be determined by the effects explaining why the mechanism producing the urgent action tendency was selected for in past circumstances of selection. According to the UMS theory of emotions, E=fear can for example be characterized as a superordinate system that controls a cluster of organismic subsystems whose synchronized operation instantiates, and manages through time, an *urgent avoidance tendency*. Such superordinate system has the pragmatic object of *danger*, which describes the conditions under which an urgent avoidance tendency is pushmi-pullyu appropriate. This is tantamount to assuming that that the system activating the urgent action tendency of avoidance characteristic of fear was selected for activating such tendency in circumstances of danger. Under this view, being elicited in dangerous circumstances is the proper function of fear.

Importantly, these may or may not be the circumstances in which an emoter is disposed to produce fear. As I will argue, the account I propose is perfectly compatible with the possibility that a given emoter may be disposed to generate urgent avoidance tendencies with respect to entirely harmless triggering circumstances. What type-identifies such tendencies as fear is that the mechanism producing them was selected for eliciting urgent avoidance tendencies when faced by danger, whether or not it currently fulfills such proper function. In other words, the history of selection of the superordinate system which activates (and manages) urgent action tendencies is what gives umotions their constitutive goals or purposes. Consequently, what matters for establishing the conditions of PP appropriateness of umotions are not the circumstances in which emoters are currently disposed to engage in urgent action tendencies, but those in which they ought to be so disposed in light of a history of selection.

To make sense of the various ingredients of this theory, and to explain how they hang together, will demand some work. I begin by clarifying the structure of the superordinate system I take an umotion to be. To do so, I need to clarify what causes the activation of an urgent action

tendency, and how its management through time is articulated into the functional components of *preparation*, *action* and *communication*.

10.2.3. Appraisal

According to the UMS theory, what causes the activation of umotions is generally the appraisal of an event.[23,24] In ordinary language, the notion of appraisal only signifies an unspecified evaluation, and there are many forms of evaluation which do not bring about umotions. What we need to understand is what sorts of appraisals are *emotional* in the sense of the UMS theory, namely such as to bring about the activation of an urgent action tendency.

The rationale for assuming that umotions are generally caused by appraisals is that (a) events of the same type can generate an umotion in some organisms but not in others, (b) events of the same type can generate different umotions in different organisms, (c) events of the same type can generate different umotions in the same organism in different circumstances. For example, events of the type "flying on an airplane" cause elation in some people, fear in others, no emotion in yet others, and different umotions at different times in the same flier (e.g. before and after a course in panic control). This suggests that it is not events in themselves – the event of flying on an airplane – that cause umotions, but rather the way in which they are evaluated by emoters.

In principle, any event can be appraised so as to cause any umotion. In practice, there is a correlation between certain types of events and certain types of umotions. Physical events such as

---

[23] I speak of "cause" in terms of Mackie's (1965) notion of cause. Mackie wrote that what is commonly – although not always - meant by saying that event A causes (caused) event B is that A is (was) "an *insufficient* but *necessary* part of a condition which is itself *unnecessary* but *sufficient* for the result" B (A is an INUS condition for B). To say that an appraisal caused an emotion is therefore to say that the *appraisal* was part, in combination with several other background conditions, of a complex condition Sufficient to bring about the emotion. The appraisal was an individually Necessary but Insufficient part of that sufficient condition, in the sense that, given the presence of the other background circumstances, the emotion would not have come about without the appraisal, and the appraisal alone would not have brought about the emotion unless the other background conditions had been present. The combination of appraisal and background conditions, finally, was Unnecessary, in the sense that the emotion could have been brought about in principle by other combinations of circumstances. There are several limitations for the INUS account as a general account of causation (see), but I will not explore them in this dissertation.

[24] The notion of an "event" is also the object of much philosophical speculation. For the purposes of this dissertation, I endorse a rough and ready view of events which ascribes to them the following key properties: (a) they are unrepeatable particulars rather than repeatable universals, (b) they occur contingently, (c) they are spatio-temporally located, (d) they have parts (see Casati and Varzi 2002). By and large, this is the view of events defended by Davidson (1980), and paradigmatically exemplified by ascriptions such as "the boiler exploded in the cellar", a concrete unrepeatable with parts that occurred at a particular spatio-temporal location. I will also use the terms *occurrence* and *episode* to refer to *events*, disregarding the differences existing between such terms in ordinary language.

loud explosions tends to elicit fear in a large variety of organisms. Mental events such as memories of past slights tend to elicit anger in human beings.

Although the UMS theory assumes that appraisals are the most common causes of umotions, exceptions are admitted. For example, there is some evidence that umotions can be chemically induced, or brought about by brain manipulations, or generated by facial feedback (Izard 1993). The evidence for such alternative forms of elicitation is not conclusive at present, but I do not see any reason to exclude that an urgency management system could be activated in ways other than through an appraisal. At the same time, I take this to be a residual case, which will not concern me from now on.[25] What makes it residual is that, as I shall argue in more detail below, umotions fulfill their proper functions when they are elicited in some circumstances rather than in others. The appraisal mechanism is what allows a given umotion to be caused in the circumstances that explain why the mechanism producing the umotion was selected for. Mechanisms of elicitation such as chemical induction, direct brain stimulation or facial feedback, on the other hand, do not represent the outcome of an evaluation of circumstances, and consequently constitute abnormal conditions of elicitation (more on this below).

Now, what are the properties of the forms of appraisal that elicit umotions? According to the UMS theory, umotion-causing appraisal has three basic properties. The first is that it is *intrinsically motivational*. If umoting is shifting to a state in which an urgent action tendency is activated and managed, umoting is being motivated to act. But since this form of appraisal cannot occur without an umotion following it (but notice: an umotion need not be preceded by an appraisal), we can speak of umotion-causing appraisal as being intrinsically motivational. This is not a discovery about appraisal, but an obvious consequence of the fact that we are considering a special class of evaluations identified as umotion-causing. Given how I characterized umotion-causing appraisal, one cannot be engaged in it without being motivated in the particular way in which an umotion motivates.

The second property is that umotion-causing appraisal is *not deliberate*, at least not in the same straightforward sense in which raising one's hand is deliberate. At least generally, one cannot "deliberate" to appraise a certain event through, say, a fear-causing appraisal rather than a disgust-causing appraisal. This aspect of emotional appraisal is one of the main reasons why

[25] Similarly, I will not be concerned with discussing whether or not we can reinterpret, say, chemically induced happiness as chemically-induced appraisal causing happiness. However we settle this issue, this is not a standard case of emotion elicitation.

emotions have been described as *passive* for most of their intellectual history. The UMS theory of emotions makes good sense of this passivity. If umotions are superordinate urgency management systems, it is not surprising that, at least in general, they cannot be elicited at will. This is because one cannot simply "will events into urgency". What the organism appraises as creating a situation of urgency is influenced by factors not available to instantaneous deliberation.

On the one hand, such factors comprise details of cognitive architecture that are beyond the reach of deliberation *in principle*. For example, there are forms of fear appraisal mediated by dedicated neural pathways which do not involve the cortical areas associated with deliberation (Le Doux 1996). On the other hand, what determines which situations generate urgency depends largely on a value system emerged throughout the entire learning history of the organism. This value system cannot be changed as easily as the position of one's arm.

This being said, we must not conflate the idea that umotion-causing appraisal is not deliberate with the idea that umotions just "happen" to emoters. This common way of thinking about the emotional is misleading in a variety of different ways. It tends to obliterate at least three dimensions of agency which importantly shape appraisal.

The first and most obvious one is that emoters can deliberately search for events they expect to appraise in a particular way. For example, one can deliberate to bungee-jump or watch a horror movie with the reliable expectation that the fear system will be activated.

The second element is that emoters can influence their value system through time. As I discussed in subsection 2.1.1, this aspect was emphasized firstly by Aristotle, who pointed out that we can influence the way we feel by *habituating* ourselves to behave in certain ways (e.g. as virtuous people do).

The third element is the one I discussed in chapter 6 under the heading of Machiavellian appraisal. As argued by Griffiths (2003), "[e]motional appraisal is sensitive to cues that predict the value to the emotional agent of responding to the situation with a particular emotion" (54). Under the UMS theory, this amounts to saying that emotional appraisal, although not deliberate, is not independent of considerations about whether or not shifting to a certain urgent action tendency will be conducive to the goals of the emoter.

This feature of appraisal accounts for the Machiavellian dimension with respect to the *activation* of an urgency management system. There is a fundamental Machiavellian dimension

also to the *management* of an urgent action tendency. This is because emoters make efforts to manage to their advantage the urgent action tendencies they undergo *throughout the emotional episode*. This Machiavellian aspect is expressed primarily through sensitivity to "affordances" predicting whether or not an ongoing urgent action tendency can or cannot be advantageously manifested (see below).

The third important property of umotion-causing appraisal is that it is *immediately followed by umotion*, in a way which is perceived as *effortless*. An emoter does not produce an appraisal first and then, with a significant time gap and the exercise of some laborious intellectual activity, an umotion follows. The fastness and effortlessness of umotion, however, must not be confused with the fastness and effortlessness of the formation of umotion-causing appraisal. Although very often appraisal is also fast and effortless upon exposure to a given stimulus pattern, on other occasions it is the culmination of a slow and effortful process. Sometimes it takes some time and effort to realize that one is in a situation requiring urgent action. Once this evaluation has been made, however, umotion follows suit.

Consider for example a pilot who notices something slightly unusual in one of his cockpit instruments. He may begin checking on several other instruments, call control tower, make some calculations, and finally realize that the airplane is unlikely to make it to the ground safely. At this point, an umotion (most likely fear) will quickly and effortlessly follow. We could debate whether or not the appraisal is not really formed until this realization is made. I see reasons to say both that what caused the umotion is the slow and effortful appraisal process which started when something unusual was noticed in the cockpit and to say that what caused it is instead the quick appraisal process instantiated by the final realization that the airplane would likely go down. I remain neutral on this issue, and allow for the possibility of describing umotion-causing appraisal as both slow and fast in such cases, provided we are clear on what we mean by what we say.

According to the UMS theory, in conclusion, umotion-causing appraisal is intrinsically motivational, non-deliberate and immediately and effortlessly followed by the umotion it causes. This being said, there are many differences between forms of umotion-causing appraisal. Firstly, different appraisals differ in terms of cognitive complexity. Secondly, different appraisals cause different umotions.

Emotion theorists have often noticed the recalcitrance of emotional appraisal to rational considerations. One may well know that airplanes are the safest means of transportation, but spend every flight in a state of terror. Facts such as these have led some theorists to assume that emotional appraisal is isolated from rational considerations, and constitutes an input system of its own. Zajonc (1980), for example, argued that "the form of experience that we came to call feeling… derives from a parallel, separate, and partly independent system for the organism" (154), which may be served by a dedicated "network in the central nervous system".

This is the idea of "affect primacy", according to which affects constitute an evolutionarily older system, independent of the system of "cognition" paradigmatically involved in activities such as categorization or recognition. My view is that this approach looks at only one side of the spectrum of emotional appraisal. Even though many episodes of emotional appraisal are indeed recalcitrant to rational considerations, other forms are not at all, or not entirely, insulated from them.

The UMS theory makes sense of this feature too. If umotion-causing appraisal registers situations calling for the pursuit of high priority goals, we must expect that different forms of appraisal will be associated with different kinds of high priority goals. In some cases, appraisal will be completely impenetrable to what one believes. For example, it is probably the case that nothing can prevent an organism from experiencing fear from sudden loss of support. At the same time, there certainly are forms of fear on which one can work by going to a psychotherapist twice a week. An example may be fear of speaking in public. In this case, the sort of emotional appraisal of a crowd which leads to an urgent avoidance tendency is at least to some degree penetrable by beliefs about one's own worth, other people's expectations and so on.

The hypothesis that there may be different levels of emotional appraisal is borne out by LeDoux's neurobiological studies on fear. LeDoux (1996) demonstrated by means of ingenious lesion studies that there is a kind of fear elicited in a reflex-like fashion through a neural *low road* that bypasses the neocortex, and projects along a subcortical pathway directly to the amygdala (see subsection 5.2.2). LeDoux persuasively argued that this neurobiological discovery indicates that "emotional responses can occur without the involvement of the higher processing systems of the brain, systems believed to be involved in thinking, reasoning, and consciousness" (LeDoux 1996, 161). At the same time, there is a *high road* to fear, which relies on cortically mediated processing systems.

The picture is most likely more complex than that, in the sense that there is no reason to suppose that there are just two, firmly distinguished levels of emotional appraisal. Most likely, there is a continuum of levels of appraisal operating at different levels of cognitive complexity. I will not attempt to further distinguish between such levels, offering just a broad characterization of what I take to be the two ends of the continuum.[26] It seems to me that such ends can be fruitfully characterized using Fodor's (1983) distinction between *modular* and *central* input systems. I propose we think of emotional appraisal as lying on a continuum such as the following:

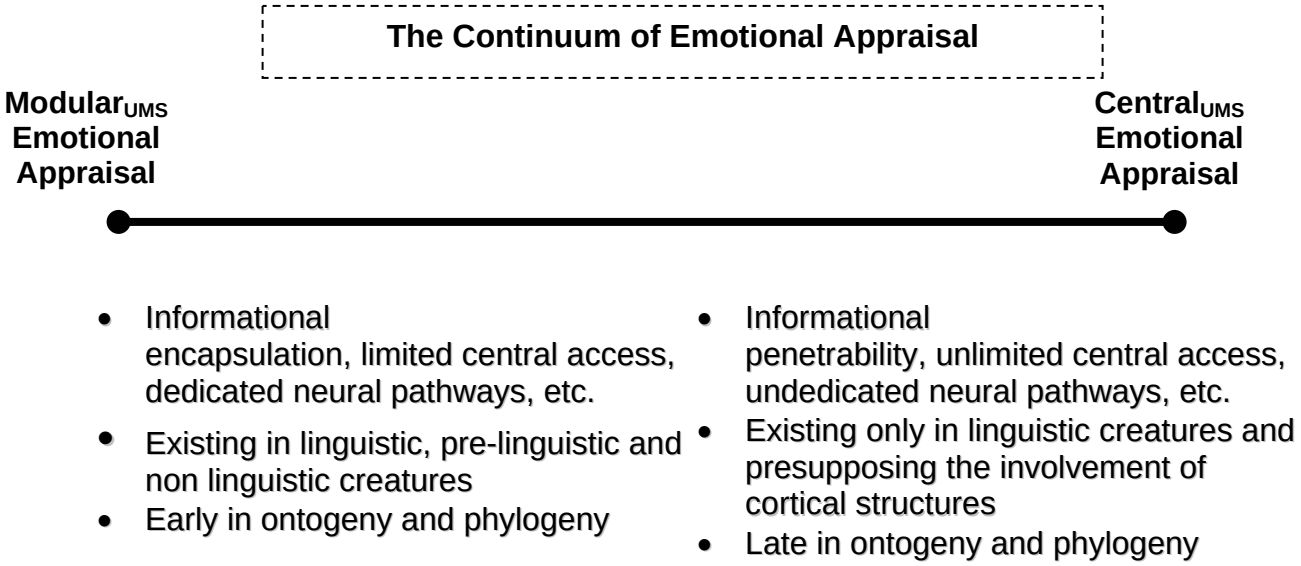


Figure 12: From modular to central emotional appraisal

As I understand them, modular$_{UMS}$ and central$_{UMS}$ emotional appraisals are characterized by *several* of the properties of Fodorian modules/central systems, but not necessarily *all* or *most* of them as required by modules and central systems *sensu* Fodor (1983). For example, I do not assume that forms of appraisal lying towards the modular$_{UMS}$ end of the continuum are innate (a trademark property of Fodorian modules). Also, I do not assume that forms of emotional appraisal lying towards the central$_{UMS}$ end of the continuum are mediated by inferences (a trademark property of Fodorian central systems). Most importantly, I take modular$_{UMS}$ forms of

[26] Leventhal and Scherer (1987) tried to distinguish between a *sensory motor level*, *a schematic level* and *a conceptual level* of emotional appraisal. The distinctions they draw have some intuitive value, but are far from being precisely characterized.

emotional appraisal to be informationally encapsulated, with limited central access and possibly dedicated neural pathways, whereas I take central$_{UMS}$ forms of appraisal to be informationally penetrable, with unlimited central access and without dedicated neural pathways.

I emphasize that I reject the strict dichotomy between modular$_{UMS}$ and central$_{UMS}$ input systems when it comes to umotion-causing appraisal. I believe that there are forms of appraisal which have some of the features of modules *and* some of the features of central systems. For example, some forms of appraisal are penetrable to beliefs, but only with major efforts. The fear-causing appraisal of a friendly crowd as dangerous, for example, may take many years of psychoanalysis to change. The idea that there are different levels of cognitive complexity in appraisal allows for the possibility that certain forms of umotion-causing appraisal are available to pre-linguistic and non-linguistic creatures, emerge early in ontogeny and are evolutionarily old, whereas others require language possession, emerge late in ontogeny and are evolutionarily more recent.

Since under the UMS theory appraisal is a detector of situations requiring urgent goal pursuit, it is to be expected that non-linguistic and pre-linguistic creatures will be able to engage in primitive forms of urgency detection. It is also to be expected that, as organisms mature, their ability to detect situations of urgency will change, and reflect the value system acquired throughout the learning history of the organism. Also, the cognitive complexity of the species to which the organism belongs will be reflected by the classes of events appraised as requiring the pursuit of high priority goals.

A second important aspect of umotion-causing appraisal, besides the fact that it lies on a continuum of cognitive complexity, is that since appraisals cause umotions and umotions differ from one another, appraisals must also differ in ways that account for the different umotions they bring about. This raises the question of what are the dimensions of evaluation to which emotional appraisal is sensitive. There is an industry in contemporary psychology devoted to characterizing the *structure* of emotional appraisal so as to determine which specific emotion will follow it. For example, Lazarus (1991, 2001) has distinguished between *primary* and *secondary* appraisals. Primary appraisals focus on answering the question “In what ways, if any, is this stimulus relevant?”, whereas secondary appraisals focus on the question “How do I cope with the situation at hand?”. Lazarus distinguished between three components of primary appraisal (goal-relevance, goal-congruence and incongruence, type of ego-involvement) and three components

of secondary appraisal (blame or credit, coping potential, future expectancy), and argued that each emotion is distinguished by a particular configuration along the six combined dimensions of primary and secondary appraisal.

Other appraisal researchers have endowed the appraisal process with even more structure. For example, Leventhal and Scherer (1987) distinguished between sixteen dimensions of appraisal, which they called *Stimulus Evaluation Checks* (SECs). In a recent update of Leventhal and Scherer (1987), Scherer (2001) collected the sixteen SECs into four classes, which represent "the major types or classes of information with respect to an object or event that an organism requires in order to prepare an adequate reaction" (94). The four classes are *appraisals of relevance, consequences, coping potential, and normative significance*. The various checks are organized by Scherer (2001, 15) into the following four classes:

1. Relevance detection checks

Novelty check, evaluating "whether there is a change in the pattern of external or internal stimulation".[27]

Intrinsic pleasantness check, evaluating "whether a stimulus event is pleasant, inducing approach tendencies, or unpleasant, inducing avoidance tendencies".

Goal/need significance check, evaluating "whether a stimulus event is relevant to important goals and needs of the organism".

2. Implication assessment checks

Causal attribution check, evaluating "the causes of the event, in particular to discern the agent that was responsible for its occurrence".

Outcome probability check, evaluating "the likelihood or certainty with which certain consequences are to be expected".

Discrepancy with expectation check, evaluating whether "the outcome is consistent with or discrepant from the state expected for this point in the goal/plan sequence".

Goal/need conduciveness check, evaluating whether the outcome is "conducive or obstructive to reaching the respective goals or satisfying the relevant needs".

Urgency check, evaluating "how urgently some kind of behavioral response is required".

3. Coping potential determination check

[27] There are three types of *novelty checks*, in terms of suddenness, familiarity and predictability.

Control check, evaluating the “degree of control over the event or its consequences”.

Power check, evaluating “the resources at [one’s] disposal to change contingencies and outcomes according to its interests”.

Adjustment check, evaluating “the potential for adjustment to the final outcome via internal restructuring”.

4. Normative significance evaluation checks

Internal standards check, evaluating whether the event “is consistent with internalizes norms or standards as part of the self concept or ideal self”.

External standards check, evaluating “whether the event, particularly an action conforms to social norms, cultural conventions, or expectations of significant others”.

For example, Scherer (2001) argues that anger is caused when an event is appraised as novel, of high significance, intentionally caused by someone else, with very high probability of consequences dissonant with the goals of the agent, requiring an urgent response, involving a high degree of coping potential, and being at odds with social norms, cultural conventions or expectations of significant others. Each emotion is similarly associated with its own profile of stimulus evaluation checks, and the holy grail of the research program is to formulate a profile that admits of no exceptions.

This approach, however, strikes me as seriously ill-conceived. One problem is that appraisal theorists seem to ascribe to emotional appraisal a high degree of cognitive sophistication, which is difficult to reconcile with many forms of emotional appraisal, especially those lying towards the modular$_{\text{UMS}}$ end of the continuum I illustrated above. Being suddenly poked in the back reliably brings about anger, but it is hard to imagine that the emotional appraisal involved in it comprises sixteen evaluation checks running in parallel in a fraction of an instant. Moreover, it is hard to imagine how pre-linguistic and non-linguistic creatures could engage in, say, *coping potential* determination checks which distinguish between appraising ”the resources at [one’s] disposal to change contingencies and outcomes according to its interests” (2001, 15) and “the potential for adjustment to the final outcome via internal restructuring” (2001, 15). Discriminations at this level of cognitive complexity definitely seem to require language possession.

A second problem is that, even if we accept the level of cognitive complexity embodied by the various dimensions of appraisal for the sake of argument, there seem to be many instances of emotion caused by evaluations different from those appraisal theorists associate with them. For example, people often get angry about events they consider of low significance, which they have themselves caused, which have very low probability of interfering with their goals, with respect to which they have no coping potential and which are not at odds with social norms, cultural conventions, or expectations of significant others. For example, I sometimes get angry with inanimate objects into which I accidentally run, even when the physical pain I suffer is minor. But I certainly do not ascribe high significance to events of this type, nor do I consider the object I run into responsible for the accident, nor do I consider my coping potential relative to an accident which already occurred particularly high, and most certainly I do not consider the event to lack conformity to social norms, cultural conventions, or expectations of significant others.

The basic problem with this literature, as I see it, is that there is a conceptual confusion at work between two notions of appraisal. One is the notion of appraisal which brings about a certain emotion (call it *emotion causing appraisal*), and the other is the notion of appraisal implied by having a certain emotion (call it *entailed appraisal*). As Kenny (1963) first argued, "each emotion is appropriate-logically, and not just morally appropriate-only to certain restricted objects" (192). As I discussed in 4.2.1, he called such objects *formal*, and assumed that each emotion is associated with a set of conditions of appropriateness that distinguish it from all other emotions. As I will argue in some detail section 10.3, the source of normativity underlying the notion of formal objects is that emotions have proper functions: they can function properly and improperly in light of their history of selection.

The formal object of an emotion E, Kenny believed, is associated with E non-contingently. In effect, the formal object describes the property complex ascribed to the contingent object of a given emotion E by having E towards it. For example, *by* fearing a given object, emoters can be said to ascribe to it the property of being dangerous. We can then distinguish between *fear-causing appraisal*, the evaluation which brings about fear, and the *entailed appraisal of fear*, the evaluation implied by having fear. Fear could be caused by direct brain stimulation rather than appraisal, but still entail a certain appraisal of the circumstances.

By distinguishing dimensions of appraisal along the lines I described, appraisal theorists have in my view simply offered a highly nuanced account of the *formal objects* of emotions.

What they have described is not what properties *emotion causing appraisals* have, but what properties *entailed appraisals* have, confusing the conditions of elicitation of an emotion E with the conditions of appropriateness of an elicited E. It is not hard to find evidence for this misunderstanding in the literature. Consider Scherer (2001) and Lazarus (1991), the two most prominent contemporary appraisal theorists. The former describes as follows the principle governing his choice of dimensions of appraisal (a.k.a. stimulus evaluation checks or SECs):

> The SECs are chosen, in a principled fashion, to represent the minimal set of dimensions or criteria that are considered necessary to account for the differentiation of the major families of emotional state (Scherer 2001, 94).

The focus here is not on what causes different emotions, but rather on what differentiates them from one another. What Scherer has tried to elucidate are in effect the basic dimensions of *entailed appraisal* which distinguish one emotion type from all others. As I read his theory, the SEC profile of an emotion describes its circumstances of appropriateness. For example, the SEC profile of anger states that anger is the sort of thing which is appropriate when elicited in circumstances of high significance, with respect to events intentionally caused by someone else, when there is a high degree of coping potential, and there is tension with social norms, cultural conventions or expectations of significant others. We may disagree with this account, but the point is that it is not an account of what causes anger, but at best of what *ought to cause* it if anger were elicited in circumstances in which it is appropriate. In turn, such circumstances depend upon the proper function of anger, which allows us to sort the causes and effects of properly functioning anger tokens from the causes and effects of non-properly functioning anger tokens.

The same problem applies to Lazarus's (1991) appraisal theory, which distinguishes six rather than sixteen dimensions of appraisal. Here is a key passage:

> I believe we should combine the partial meanings, which derive from a causal analysis of a number of part processes-that is, the appraisal components, of which I have enumerated six-into a terse, integrated gestalt or whole, which is what characterizes the cognitive-motivational-relational cause of the emotion. In other words, the process of appraising must be examined at higher level

> of abstraction that just a listing of separate, partial meanings. I refer to this higher level as the core relational theme for each emotion. This theme is a terse synthesis of the separate appraisal components into a complex, meaning-centered whole (Lazarus 1991, 64).

Lazarus (1991) suggests that the core relational theme of sadness is having experienced an irrevocable loss, the core relational theme of anger is having experienced a demeaning offense against me or mine, the core relational theme of fear is having experienced immediate, concrete and overwhelming physical danger, the core relational theme of guilt is having transgressed a moral imperative, and so on. Lazarus' core relational themes, once again, correspond by and large to the formal objects of emotions. This is not surprising, because they are not descriptions of the "cognitive-motivational-relational cause of the emotion", but rather of the "restricted object" an emotion must be associated with in order to be what it is.

People can be angry without having experienced a demeaning offense (e.g. anger towards inanimate objects), afraid about things they do not appraise as representing immediate, concrete and overwhelming physical dangers (e.g. panic attacks), guilty about events which they do not believe to involve the transgression of a moral imperative on their part (e.g. victim guilt), and so on. If we take core relational themes to be good descriptors of conditions of appropriateness, these forms of anger, fear and guilt qualify as *inappropriate*, because they violate the constitutive norms of appropriateness of the emotions they are. In other words, they represent tokens of emotions that fail to fulfill their proper functions.

Every core relational theme can be broken down into smaller conceptual components. A demeaning offense against me or mine, for example, can be individuated on a hyperspace with six dimensions – the one Lazarus proposes - by claiming that it amounts to the entailed appraisals that something relevant to my well-being happened, that it is bad, that I or somebody I care about is involved, and that someone is to blame for it. But it is misleading to say that these dimensions, which are *conceptually related* to being an offense against me or mine, are combined into "a terse, integrated gestalt or whole" which *causes* anger.

The study of emotion causing appraisal, I suggest, must explore empirically what kinds of situations generate what kinds of emotions. Appraisal theorists, however, rely for the most part on self-reports, as they try to figure out what caused an emotion by asking people what goes through their heads when they experience it. But, as insightfully argued by Parkinson (2004,

117), “[w]hen people usually think about their emotions, the default option is to think of justifiable, reasonable examples as specified in everyday common sense”. This is another way to put my point, which is that the dimensions of appraisal obtained by appraisal theorists – generally through self-reports - are dimensions of appropriateness, which describe (at best) the etiology of *justified emotions*. Let us now turn from the activation of emotion to its management, and try to understand how exactly an urgency management system works.

10.2.4. From appraisal to preparation, action and communication

In this section, I aim to clarify the following portion of the central thesis of the UMS theory of emotions:

> An umotion E is a superordinate system, generally activated by an appraisal, which controls a cluster of organismic subsystems whose synchronized operation instantiates, and manages through time, an action tendency T_E with control precedence.

This is not a complete account of umotions, because no mention is made of their conditions of appropriateness, to be discussed in the next section. My task is to clarify how the managing operations of the superordinate system “umotion” work, illustrating the three functional components I called *preparation*, *action*, and *communication*. The following picture can help us organize our discussion:

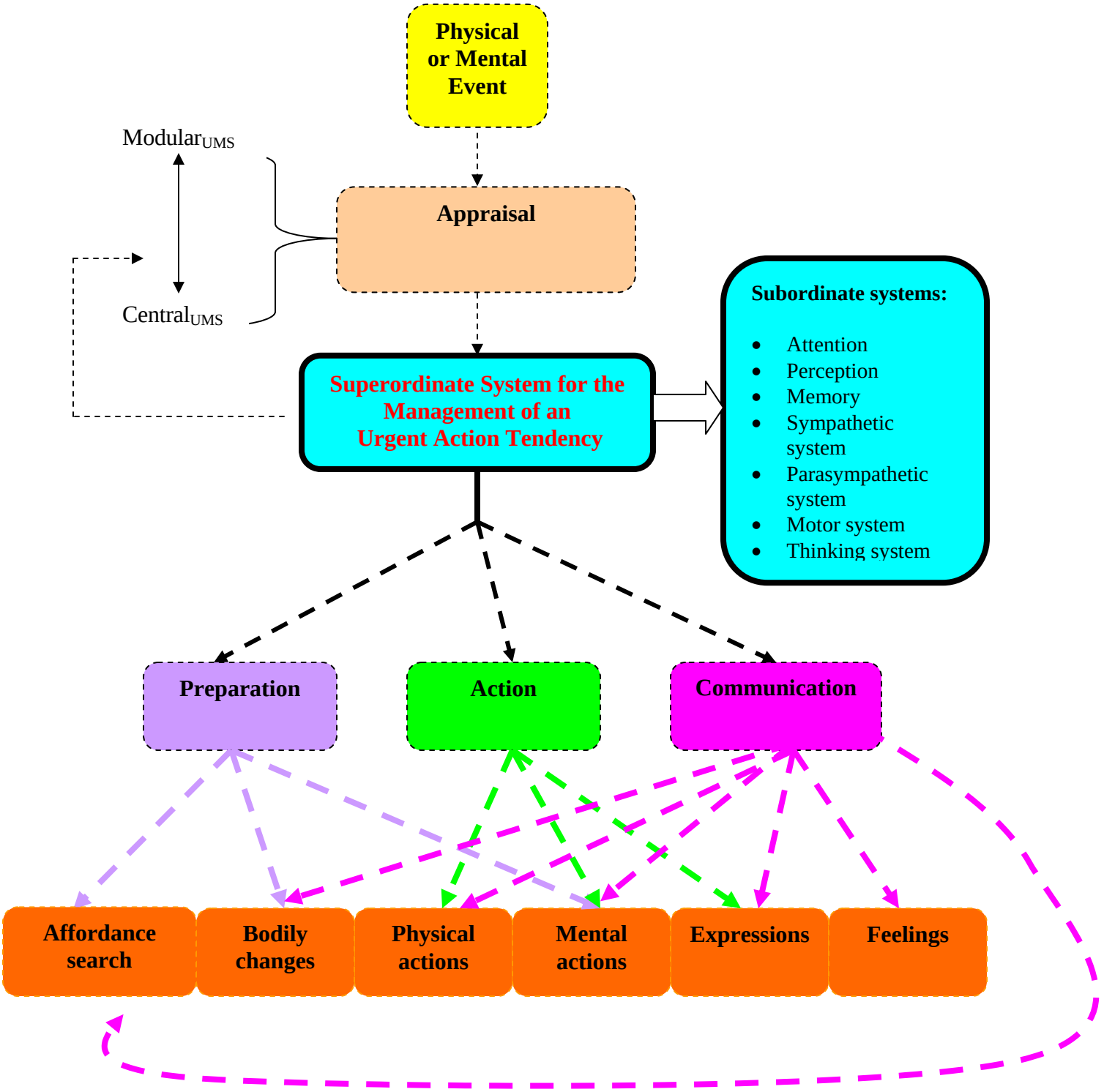


Figure 13: A diagram of umotions as urgency management systems

Dotted lines represent relations instantiated in many or most instances of umotion, but not strictly necessary for umotion to be instantiated. Non-dotted lines represent instead relations which are necessary to have umotion. Here is, in broad outline, what the diagram is meant to represent:

(a) Umotions are generally caused by appraisals of physical or mental events, but exceptions are admitted (e.g. facial feedback, direct brain stimulation)

(b) An umotion is activated when a superordinate system for the management of an urgent action tendency is activated

(c) The superordinate system that is umotion controls a variety of subordinate subsystems such as attention, perception, memory, the parasympathetic system, the motor system, and so on in the specific ways involved in the management of a specific urgent action tendency

(d) The management of an urgent action tendency has two main dimensions, one *tendency-focused* and the other *goal-focused*

(e) An important component of the *tendency-focused* management is the re-appraisal of the eliciting event, portrayed by the dotted line between the superordinate system and appraisal. For example, a snake-shaped object in the dark may automatically activate the fear system through a $\text{modular}_{\text{UMS}}$ appraisal, and then be re-appraised and judged not to be dangerous. This would most likely lead to the inhibition of the action tendency. As shown in the diagram, re-appraisal tends towards the $\text{central}_{\text{UMS}}$ end of the continuum of emotional appraisal. The snake-shaped object will be categorized as being or not being a snake, and beliefs about the particular snake it is will be involved in re-appraising whether or not it is dangerous.

(f) The *goal-focused* management of an urgent action tendency comprises three main functional components, namely *preparation*, *action* and *communication*. Without the instantiation of at least one of them, there is no umotion, hence the full line below the superordinate system box. This full line divides up into three dotted lines, which are meant to indicate that an umotion can be instantiated by preparation only, by action only and by communication only. In prototypical cases of umotion, all three functional components will be involved.

(g) The *preparation* component of umotion has to do with the fact that, once an urgent action tendency is activated, emoters search for ways to fulfill its relational goal (affordance search), undergo bodily changes (e.g. increased heart rate), and engage in mental actions (e.g. planning ways to respond to possible moves of a predator). Some umotions are so fast, however, that they lead to action directly, without going through any preparation.

(h) The *action* component of umotion has to do with the fact that, once an action tendency is activated and opportunities to manifest it advantageously have been found, emoters try to fulfill the relational goal of the action tendency. Their actions can be physical actions, mental actions or expressions. Some emotions, however, do not lead to action but rather to *inhibition*. This is often because no affordances for an advantageous manifestation of the action tendency have been recovered.

(l) The *communication* component of umotion has to do with the fact that, once an action tendency is activated, emoters broadcast signals non-arbitrarily associated with the urgent action tendency that is being managed. Some signals are directed to the self. For example, feeling that one's body is trembling may work as a signal to the self that one is undergoing fear. Some signals are instead sent to others. For example, one can send signals of anger through facial expressions (e.g. a fixed stare) or through physical actions (e.g. slamming a door). Importantly, *communication* has an impact on *preparation*: responses to emotional signals are factored in the search for ways to fulfill the relational goal of the action tendency. For example, the responses of an interactant to an anger expression will be factored in the evaluation of whether or not the interactant is of a type which allows for anger to be manifested advantageously through angry physical actions. Some umotions are so fast, on the other hand, that they lead to action directly, without going through any communication of signals prior to taking action.

These are the basics of the theory I will articulate and defend. Let us now get into the details.

10.2.5. Umotion as a superordinate system

I claimed that an umotion is a superordinate system which controls a cluster of organismic subsystems whose synchronized operation instantiates, and manages through time, an urgent action tendency. The term "system" comes from the Greek "synistanai", which means "to combine". By using it, I intend to emphasize that an umotion is a complex entity comprising interrelated parts which form a unified whole. What makes an umotion *superordinate* relative to other *subordinate* systems is that it controls them in a coordinated fashion. Although such subordinate systems, in turn complex collections of integrated parts, exist independently of umotion, the synchronized way in which umotion can recruit them suggests that they can be conceived as parts of a higher order entity that is the umotion.

However, they are not subordinate systems in the sense that their only purpose is being part of an umotion, as the purpose of a heart ventricle is being part of the heart. Rather, such systems fulfill other functions unrelated to umotion. For example, the motor system supports non-emotionally navigating a cluttered environment as well as engaging in the evasive actions of fear.

An important feature of the theory I propose is that umotion controls a "cluster" of subsystems. The notion of a cluster is meant to signal that there is no specific set of subsystems *always* recruited when an umotion is instantiated. Various combinations of subsystems can instantiate and manage any given urgent action tendency type, and different action tendencies will recruit different groups of subsystems. The type of umotion at hand, and the specific circumstances of its elicitation, will determine which systems are recruited in a given episode of umotion.

Now, where does the idea that emotions are superordinate systems come from? Although several emotion theorists have described the emotions as systems (e.g. Simon 1967), and suggested that they coordinate the operation of several subsystems (e.g. Scherer 2001), my main inspiration comes from Tooby and Cosmides (1990, 2000). An important difference between my theory and theirs, however, is that they take an emotion to be a *superordinate program evolved to deal with ancestral life tasks*, whereas I take an umotion to be a *superordinate system which instantiates and manage a state of action readiness endowed with control precedence.* I leave the question of the origin of such superordinate system open for discussion, allowing for the possibility that some umotions did not evolve to deal with ancestral life tasks. Consider the following passage from Tooby and Cosmides (2000):

> [A]n emotion is a superordinate program whose function is to direct the activities and interactions of the subprograms governing perception; attention; inference; learning; memory; goal choice; motivational priorities; categorization and conceptual frameworks; physiological reactions…; reflexes; behavioral decision rules; motor systems; communication processes; energy level and effort allocation; affective coloration of events and stimuli; recalibration of probability estimates, situation assessments, values, and regulatory variables (e.g., self-esteem, estimations of relative formidability, relative value of alternative goal states, efficacy discount rate); and so on (93)

This list comprises redundancies (e.g. goal choice and motivational priorities), it runs the risk of circularity (e.g. affective coloration of events), it characterizes in a coarse-grained fashion some of the systems involved (e.g. it treats all physiological reactions as if they were part of one system), and it is not exhaustive (e.g. the musculoskeletal response system is not included). Consequently, I do not endorse the list it in its current form, but borrow a key insight from it. This is that an umotion has *global reach* over *practically all* cognitive, perceptual and motoric resources available to the organism. There is no need to fuss over what specific subsystems are recruitable by emotion, once we realize that in circumstances of urgency practically all of them are. What is missing from Tooby and Cosmides's account is something much more important, namely an explanation of the way in which the resources of the organism get to be coordinated by the superordinate system.

Tooby and Cosmides (2000) have in effect offered a laundry list of subsystems without explaining their role in the superordinate system that is the emotion. Their account suggests that a superordinate system counts as an emotion just because it directs "the activities and interactions" of a bunch of subsystems. As they put it, an emotion "is a superordinate program whose function is to direct the activities and interactions of the subprograms". I see two main problems with this account.

The first is that emotions are not the only systems capable of directing the activities and interactions of several subsystems. For example, the "system" which governs navigation through a cluttered environment controls the activities and interactions of attention, goal choice, reflexes, motor systems, effort allocation and so on, but it is not an emotion. The second problem is that the proper function of emotions is not merely to coordinate subsystems, but to obtain something by means of such coordination. What is missing from Tooby and Cosmides's account is the

central idea that an emotion directs the activities and interactions of subsystems *in pursuit of high priority goals.*

Failure to bring this point to the fore may be due to the assumption that emotions evolved to deal with ancestral life tasks. As I discussed in 5.2.1, ancestral life tasks are characterized by Tooby and Cosmides (2000) as having five main characteristics: they recurred ancestrally, they could not be successfully negotiated in the absence of a superordinate level of program coordination, they had a rich and reliably repeated structure, they had reliable cues signaling their presence, and they were of a type in which an error would have resulted in large fitness costs. Tooby and Cosmides may be convinced that once the idea of evolutionary origin is added to the idea of a superordinate program, what we get is a good principle of individuation for emotions. But this is far from being the case.

Firstly, the five characteristics fail to distinguish between emotions and other superordinate systems of possible evolutionary origin. For example, the superordinate systems for navigating a cluttered environment may arguably have evolved to deal with situations that recurred ancestrally, required a superordinate level of program coordination, had a rich and reliably repeated structure and reliable indicators, and created the risk of large fitness costs (e.g. hitting an obstacle at high speed). However, a navigation system is not an emotion.

Secondly, there is no reason to assume that all emotions evolved to deal with situations with the features described. Evolutionary psychologists are generally convinced that the mind is "a crowded zoo of evolved, domain-specific programs" (Tooby and Cosmides 1990, 91), and they apply this assumption to emotions. As I argued in 5.2.1, the evidence on which they rely is often very thin, and it has led to the mistaken expectation that emotions are highly domain-specific programs, whereas there are good reasons to expect them to be open programs in Ernest Mayr's sense (Griffiths 1997; see section 5.2). To mark the distance between my account and the one proposed by evolutionary psychologists, I call emotions "superordinate systems" rather than "superordinate programs", even though "program" and "system" are to a large extent semantically equivalent.

Progress with respect to Tooby and Cosmides' (2000) account demands not only that we avoid making their mistakes, but also that we fill in their omissions. What we need to understand is how the superordinate system that is umotion organizes subsystems so as to instantiate and manage an action tendency with control precedence.

10.2.6. Preparation: body and mind

According to many emotion theorists, the most important function of emotions lies in *bodily preparation*. Descartes (1650), for example, argued that "the principal effect of all the human passions is that they move and dispose the soul to want the things for which they prepare the body" (art. 40) (see chapter 2). We can distinguish various types of bodily changes, which will be involved in different types of urgent action tendencies. [28]

The bodily changes most often invoked in emotion theory are autonomic changes, namely the sorts of changes James (1884) presupposed when he claimed that an emotion is a perception of bodily changes. *Autonomic changes* comprise changes in heart rate, blood pressure and blow flow distribution, respiration, electrodermal activity and sweating, gastrointestinal and urinary activity, secretory and papillary responses, composition of blood and saliva and trembling. But there are also other kinds of changes which may in principle count as bodily.

Three further examples are *hormonal changes* such as changes in the catecholamine hormones epinephrine and norepinephrine, *musculoskeletal changes* such as changes in muscle tension, and *neural changes* such as changes in the activation of particular brain areas (e.g. amygdala activation in fear).

In the context of the UMS theory, physiological responses are understood in terms of the role they play in the activation and management of an urgent action tendency. Generally speaking, bodily preparation is preparation for the execution of the kinds of bodily movements that are liable to fulfill the relational goal of the action tendency. Since different action tendencies have different relational goals, there may be some degree of emotion specificity in bodily changes. In section 7.1.1, I concluded on the basis of the review of the available empirical evidence offered by Cacioppo et al. (2000) that there are some broad differences between families of emotions (e.g. positive and negative emotions) and some specificities with respect to some emotions (e.g. heart rate increase in intense fear is generally higher than in any other emotion), but nothing like a one-to-one correspondence between emotion types and bodily profiles.

[28] The distinctions between kinds of bodily changes, and the lists of specific bodily changes, are borrowed from Frijda (1986).

Emotional preparation has also a *mental* counterpart. The subsystems governing physiology are not the only ones to be mobilized when emoters prepare for action. Consider a gazelle drinking by a pond. A predator enters her visual field, appraisal follows, and the fear system is activated. What does this involve? On the one hand, subordinate systems governing physiological reactions will be activated. There will be an increase in cardiac output, breathing rate, and diastolic blood pressure, enhanced skin conductance responses and gastrointestinal activity, activation of the amygdala and so on.

On the other hand, the systems governing faculties like attention, perception, memory, etc. will also be recruited and put at the service of the management of the urgent avoidance tendency. For example, the attention of the gazelle will be redirected from drinking to dealing with the presence of the predator. Her perceptual abilities will be sharpened, so that she can see, hear, smell, touch and taste with maximal acuity. As suggested by Tooby and Cosmides (2000), an animal in the circumstances described may also activate a "specialized inference system" according to which a predator's "trajectory or eye direction may be fed into systems for inferring whether the [predator] saw you" (94). The memory system would also be activated. The gazelle may suddenly remember that she just walked by a pathway which may lead her to safety. This conclusion may be facilitated by "conceptual frame shifts", as the gazelle may apply "safety categorization frames" (104) to what she perceives, infers or remembers. The list of mental activities does not have any principled limitation other than the cognitive complexity of the organism under consideration (e.g. a human being may engage in *imaginings*, *intendings,* and so on).

The general point is that not only the body but also the mind of the emoter is *readied* in the course of the management of an urgent action tendency, and put at the service of the goal pursuit that characterizes it. What is rarely discussed in emotion theory is the active side of emotional preparation, namely the search for means to fulfill the relational goal of the action tendency. Given its importance and relative neglect in the contemporary literature, I will now say a little more about what I call the *affordance search*.

10.2.7. Searching for relational goal-affordances

There is a picture of emotions which has loomed large in the history of the subject. According to such picture, emotions just happen to emoters when they are presented with the appropriate stimuli. This picture is not entirely off the mark. In some cases, umotions work indeed as reflexes. You present an organism with a sudden and unexpected loud noise, and the fear system will be reliably activated. In many cases of umotion, however, this picture is very misleading. What it hides from view is the fact that an umotion comprises several active components, one of which is the *search for opportunities to fulfill the relational goal of the action tendency*. I already mentioned that Machiavellian considerations enter already at the level of emotional appraisal, since the organism is sensitive to cues predicting whether or not a certain emotion could possibly be to his or her advantage (Griffiths 2003). What I have not yet discussed is what shape such considerations take *after* an urgent action tendency has been elicited.

The way I propose to make sense of the active side of emotions is through the notion of affordances. The notion of affordances was introduced in psychology by James J. Gibson (1979). It is an important concept, which has not yet received the attention it deserves (but see Bermudez 1998, Machamer and Osbeck 2003, and especially Millikan 2004). In part, this is because affordances have been associated with some of the problematic assumptions of the Gibsonian movement. Most importantly, they have been associated with Gibson's rejection of the notion of representation as a useful explanatory construct. This position puts Gibsonians at loggerheads with most cognitive scientists working today.

As I argued in Scarantino (2004), however, the concept of an affordance can be made sense of independently of the thesis that perception is *direct*. This is because whether or not X affords Y to O, where X is a portion of the environment, Y is an action mode and O is an organism, does not depend on whether or not O perceives that X affords Y. At the same time, it must be emphasized that the theoretical interest of the notion of affordances is tied to their being perceivable.[29] Gibson believed that a great many affordances are specified in ambient energy and that as a matter of empirical fact organisms guide their behaviors by becoming attuned to them.

Gibson introduced the notion of affordances to capture the essential complementarity between the organism and the environment. As he put it, the "*affordances* of the environment are

[29] I thank Ruth Millikan for urging me to emphasize that the theoretical interest of affordances is fundamentally tied to their perceivability

what it *offers* the animal, what it *provides* or *furnishes*, either for good or ill" (Gibson 1979, 127; emphasis in original). For example, Gibson described a surface such as the brink of a cliff as *fall-off-able*, a substance such as an apple as *eat-able*, an object such as a stone as *throw-able*, an animal such as a conspecific as *copulate-with-able*, an event such as a fire as *cook-with-able*.

The two key dimensions of affordances are *relationality* and *potentiality*. *Relationality* is due to the fact that the offerings of the environment cannot be specified independently of the organism (or class of organisms) relative to which they are instantiated. A lioness offers an opportunity to copulate to a lion, but certainly not to a mouse, to a gazelle or to an elephant.

Potentiality is due to the fact that the offerings of the environment are contingent upon what *may* be the case, were some further circumstances to occur. The fact that a lioness affords copulation to a lion does not imply that copulation will occur, but only that copulation is possible given certain circumstances (e.g. the lion is interested in it). In Scarantino (2004), I tried to capture the relationality and potentiality of affordances by assimilating them to particular kinds of dispositional properties. The basic idea I defended is that to ascribe an affordance to some bearer X relative to some organism O in certain background circumstances is to say that, in such circumstances, if a set of triggering circumstances T were the case, then a manifestation M involving X and O would be the case (with significant probability).

For example, to say that a given tree affords *climbing* to a squirrel in normal ecological circumstances (e.g. the tree is not covered by an invisible slipping substance) is to say that, in such circumstances, if the squirrel were to try climbing, he would be highly likely to succeed. Under the Gibsonian picture, the perception of a climbing affordance plays a causal role in allowing successful climbing to occur. The squirrel not only perceives that the tree affords climbing, but also perceives how to climb, namely what specific dynamic sequence of his own movements affords climbing.[30]

In Scarantino (2004), I distinguished between two main types of manifestations of affordances. If we look at Gibson's examples of affordances, we notice manifestations such as climbing, catching, getting under, eating, mailing a letter, but also such as bumping into, getting burned by, falling off, being eaten by. Whereas events in the first list constitute things organisms *do*, events in the second list constitute things that *happen* to them. The distinction is in my view

[30] I owe the clarification of this point to Ruth Millikan's comments on an earlier draft

important enough to distinguish between two classes of affordances, namely *goal-affordances* (their manifestation is a *doing*) and *happening-affordances* (their manifestation is a *happening*).

Now, Gibson assumed that perception is always *perception of affordances*, namely of possibilities for actions (and for happenings, which in turn influence the organism's possibilities for actions):

> [P]laces, attached objects, objects, and substances are what are mainly perceived, together with events, which are changes of these things. To [perceive] these things is to perceive what they afford. (Gibson 1979, 240)

The crucial *empirical hypothesis* of ecological psychology is that *affordances are perceivable*. At the same time, "an affordance is not bestowed upon an object by a need of an observer and his act of perceiving it" (Gibson 1979, 139). It is bestowed instead by a set of *physical* properties of the affordance-bearer and the organism which are relevant to make a specific activity possible (e.g. grasping, catching, being eaten by). For example, Gibson (1979, 133) indicated that "to be graspable, an object must have opposite surfaces separated by a distance less than the span of the hand."

But how do organisms perceive affordances? To perceive affordances, according to Gibson, is to *become attuned* to *invariants* and *disturbances* that *specify* them. An intuitive understanding of these technical notions is the following. An *invariant* is a property of the structure of ambient energy arrays[31] (e.g. the optic array, the acoustic array, etc.) instantiated when, relative to some source of change such as a moving point of observation or a moving source of illumination, the structure is left *unchanged* in a way that is *typical* of the item specified (e.g. a reflectance can specify the *substance* "coal" by being *unchanging* in the way characteristic of coal substances). A *disturbance* is a property of the structure of ambient energy arrays instantiated when, relative to some source of change (e.g. the change constituted by an approaching predator), the structure presents a *pattern of change* that is *typical* of the item specified (e.g. the contour of an animal can specify the *event* "approaching predator" by *changing* in the way typical of approaching predators).

[31] "To be an *array* means to have an arrangement, and to be *ambient at a point* means to surround a position in the environment that could be occupied by an observer" (Gibson 1979, 65).

In general terms, to say that affordances are *perceivable* is to say that *there are* invariants and disturbances in ambient energy arrays that specify the threats and promises of items in the environment. For example, to say that the eat-ability of a given apple is perceivable, or that the being-hit-by-ability of a flying ball is perceivable is to say that there is a sensory appearance - a way to be visible/audible/tangible/odorous/tastable - typical respectively of apples affording eating, and of flying balls affording being hit by. Gibson was very clear that we cannot establish "a priori" what affordances are specified in ambient energy.[32] As he put it, "[t]he central question for the theory of affordances is not whether they exist and are real but whether information is available in ambient light for perceiving them." (Gibson 1979, 140) Information is available for perceiving all and only those offerings of the environment that are associated with typical sensory appearances. In some cases, the organism will have to *learn to perceive* a perceivable affordance, i.e. learn to become attuned to the invariant or disturbance specifying it. What invariants and disturbances are in fact available for the specification of affordances will have to be established by empirical investigation.

But which of the innumerable perceivable affordances will in fact be perceived by an organism? As Millikan (2004, 164) persuasively remarks, "[t]here seems no reason to suppose that affordances irrelevant to current needs are always, or even ever, perceived by most animals". Generally speaking, not all affordances of the environment can be perceived at any one time.

There are innumerable things one can do, and innumerable things that can happen: some principle of selection as to what affordances are in fact perceived at any one time is necessary. For example, my red pen affords being brought to Rue Moliere in Paris in 2007 at 3:32 pm, and my neighbor's dog Terry affords being roasted together with an orange and the finger of a Congressman from Iowa. To suppose that anytime I look at the pen or at Terry I perceive such affordances would clearly be preposterous. Which among the affordances of my red pen or Terry do I perceive at any one time? Millikan's suggestion goes exactly in the right direction: "[m]ore likely [organisms] only perceive what they have motivation, at the moment, to exploit"` (164). The intuitive appeal of this idea is nowhere more evident than in the case of umotion. The activation of an umotion generates a principle of selection for affordances, according to which

[32] Gibson (1979)'s main focus was on ambient *light*, but his account must be generalized, because an affordance could be specified *despite* lacking a typical way to *look* (it may have a *typical* way to, say, *sound* and/or *smell*).

organisms perceive what can help them fulfill the relational goal of the urgent action tendency associated with umotion.

Consider once again our gazelle placidly drinking by a pond, a few instants before she sees a predator. The environment affords her all sorts of things, some good and some bad. For example, the pond affords drinking, but also drowning in. A tree perched on a cliff in the distance affords eating berries, but also presents the danger of falling off the cliff. A large passage between two trees affords running through, whereas a narrow passage affords bumping into the neighboring trees, and so on. The appearance of the predator activates a superordinate system whose management encompasses, among other things, a prioritized search for "escape the predator"-affordances.

This will entail the interruption of all prior affordance searches (what portion of the terrain affords drinking without falling into the pond?) and of all ongoing actions (e.g. drinking), and a global redirection of organismic resources towards searching for means to successfully avoid the predator.

One key aspect of emotional urgency, therefore, is that the superordinate system umotion puts organismic subsystems at the service of a *relational goal-affordance search*. In the case of the gazelle, this will be a search for affordances that fulfill the relational goal of avoiding the predator. I speak of "searching" rather than "perceiving" relational goal-affordances because I do not assume that the only faculty mobilized to recover affordances is perception. The subsystems governing what Tooby and Cosmides (2000) call Memory and Inference, just to give two examples, may be essential to discover what affords escaping a predator.

The idea of an *affordance search* allows us to bring into relief the fact that umotions, at least when not operating reflex-like, are peculiar action control structures combining *urgency* with *flexibility*. Urgency is due to the fact that all organismic subsystems are coordinated towards a search for relational goal affordances, and the exploitation of such affordances when available. Flexibility derives from the fact that an affordance search may unveil several goal affordances for the same relational goal, which can lead to a variety of different actions. At the same time, the affordance search may indicate that there simply are no available relational goal affordances in the circumstances.

Importantly, what goal-affordances are available changes dynamically through time, often because the behaviors of other organisms change as the emotional episode unfolds. One of the

main channels through which the goal-affordances space is dynamically shaped is the communication of emotional signals to other organisms. I will come back to this point shortly, after having illustrated the *action* component of umotion, which comprises emotional expressions, one of the main channels of emotional communication.

10.2.8. Action: physical, mental and expressive behaviors

The notion of action is notoriously hard to pin down, as soon as one tries to go beyond the platitude that an organism's actions comprise whatever the organism *does*. For the purposes of this dissertation, I take an action to be an *event that is goal-directed under some description*, namely caused by the pursuit of a goal and such that there is a description of the event under which the goal is achieved (in the right way). By emotional *action*, I mean anything the organism *does* with the aim of fulfilling the *relational goal of the action tendency*.

This notion of action must not be confused with the notion of action prevalent in the philosophy of action. According to the widely influential Causal Theory of Action, for example, acting is doing something with an intention, and "[i]f someone acts with an intention then he must have attitudes and beliefs, from which had he been aware of them and had he had the time, he could have reasoned that his act was desirable" (Davidson 1980). More precisely, an action is understood as an event that is *intentional under some description*, namely triggered by an intention and such as to fulfill it under some description (in the right way) (Davidson 1980, 61). By speaking of emotional action as being goal-directed under some description I mean to resist two equally detrimental tendencies in the study of emotion.

The first is that of assuming that emoting is shifting from a domain in which goals are pursued to a domain in which things just happen to emoters. This is a view that first emerged with the Stoic account of the passions as "excessive impulses which are disobedient to reason" (Arius Didymus, 65 BC, as quoted in Baltzly 2004), and it has loomed large in theories of the emotions ever since (see my discussion of appraisal in 10.2.3). According to the UMS theory, instead, emotions have a crucial goal-directed aspect, in the strong sense that they are special systems for the pursuit of goals, paradigmatically characterized by a combination of urgency and flexibility.

The second bad habit in emotion theory has been to assume that emotional action can be assimilated to the standard model of action presupposes by the Causal Theory of Action I

sketched above. Under this view, an action is always understood as the culmination of an actual or "as if" process of practical reasoning. As a first approximation, practical reasoning consists of constructing a space of alternatives, predicting the consequences of each and their probability of occurrence, weighing pros and cons associated with each alternative in light of one's beliefs and desires, and finally forming an all-things-considered intention to select one of the alternatives, and executing such intention when the time to do so has come.

The problem with this approach is that emotional action is not brought about by an actual process of practical reasoning, but by a superordinate system for the management of an urgent action tendency. As I argued above, such system is generally activated by an appraisal which is intrinsically motivational, non-deliberate and immediately and effortlessly followed by the activation of the superordinate system umotion. It is highly misleading to assimilate the operations of appraisal and of umotion, which jointly lead to emotional action, to practical reasoning. This view misses entirely what is special about umotion as an action control structure.

One may try to solve this difficulty by assimilating appraisal to "as if" practical reasoning, as philosophers have done to explain automatic actions such as standing up from a couch to get a beer while thinking about something else. The problem with this move is that it fosters the mistaken expectation that the etiology of emotions is grounded upon one's beliefs and desires, in the same way in which automatically standing up to get a beer may be. Although the analogy may work with some cases of appraisal, it is patently false when appraisal lies towards the $\text{modular}_{\text{UMS}}$ end of the continuum I described in section 10.2.3.

Forms of $\text{modular}_{\text{UMS}}$ appraisal tend to be shared across species, appear early in ontogeny, and are endowed with properties such as informational encapsulation, limited central access, dedicated neural pathways, and so on. These properties are not compatible with the view that emotional appraisal is the culmination of an "as if" practical inference, unless we posit special classes of emotional beliefs and desires which have exactly the characteristics we need to make sense of the features of $\text{modular}_{\text{UMS}}$ emotional appraisal.

But this would be an ad hoc move, without any foreseeable explanatory purchase. If all we mean by saying that $\text{modular}_{\text{UMS}}$ appraisal is an "as if" practical inference is that it causes action in the specific way in which appraisal does it when it causes umotion, then we are not making any progress in our understanding of the phenomena, but just redescribing them in a potentially misleading way. What makes it misleading is that many philosophers make assumptions about

beliefs and desires - e.g. that they require language possession, that they can be recombined with maximal inferential promiscuity (Hurley 2003), that they comprise smaller units such as concepts, etc. - which are squarely at odds with the properties that would have to be posited for them to make sense of modular$_{UMS}$ forms of appraisal. The likely result would be that holders of different views of beliefs and desires would begin talking at cross purposes, which is exactly what has happened with respect to the idiosyncratic cognitivist understanding of the notion of "judgment" (see my analysis in chapter 7).

A characteristic feature of the UMS theory of emotions I propose is that emotional actions are considered to be a *part* of umotion. In emotion theory, instead, actions are generally considered to be caused by emotions, but not part of them. Notably, this is true even of Frijda's (1986) theory, where emotions are described as action tendencies with control precedence. As far as Frijda's theory is concerned, an emotion *ends* with the emergence of an urgent action tendency. Under the view I propose, instead, the management of an urgent action tendency *is* what the umotion system is all about, and a key aspect of such management is the transformation of the action tendency into an actual action.

I find it paradoxical that so many emotion theorists have considered bodily changes, which amount to preparation for action, as being essential to emotion, but not the action to which they prepare, as if the point of emoting were simply to prepare for doing things without actually doing them.

One reason why action has been kept distinct from emotion may be that not all emotions appear to lead to actions. According to the UMS theory, the superordinate system that is umotion leads to *action* whenever it leads to facial, vocal and postural *expressions*, whenever it leads to *physical actions* involving motor control (e.g. moving, maintaining posture), and whenever it leads to *mental actions* such as thinking, reasoning, planning, reminiscing, etc. This is because these are all events which are goal-directed under some description. The distinction between expressions, physical actions and mental actions is hard to draw precisely, and I will only be able to rely on an intuitive understanding of it, which admits of borderline cases. For example, weeping appears to be in some sense an expression and in some sense a physical action, and talking appears to be in some sense a physical action and in some sense a mental action.

Even with this broad understanding of action, I agree that not all emotions lead to action. An alternative is that the superordinate system managing the urgent action tendency leads to *inhibition*. The inhibition of an urgent action tendency amounts to its active suppression, which is something one *does* in an inverted commas sense at best. Inhibition can be brought about in a variety of different ways. One may be through the operation of tendency-focused management, as when emoters re-appraise the event that activated umotion and realize that the circumstances do not justify the elicited urgent action tendency (e.g. the snake-shaped object is a toy). Another may be through change of the eliciting circumstances, as when the event that caused and justified the urgent action tendency goes out of existence (e.g. a real snake is shot by someone else). On other occasions, inhibition comes about when the search for affordances leads to the conclusion that there are no opportunities to fulfill the relational goal of the action tendency.

The possibility of inhibition points us to two distinct aspects of the potentiality of urgent action tendencies. On the one hand, they can be actively suppressed. On the other hand, there is a variety of ways in which they can be behaviorally manifested. This leads us to another possible worry, which may also explain why emotion theorists have generally taken emotional action to be distinct from emotion. Given any umotion, there appear to be innumerable forms of action to which it can lead, especially if we allow action to range over physical action, mental actions and expressions. This is true, but it is also true that the set of actions one can perform when in a state of urgent action tendency is not arbitrary.

To "umote", under the view I propose, is to have an urgent tendency to do any of the innumerable actions *which fulfill a given relational goal*. The avoidance of a danger can occur in a fearful person by running away, by staying put, by taking cover, by calling for help, by brandishing a weapon, and so on. These are all means to the same higher order goal, namely that of avoiding the danger. Trying to describe the relational goal of a given urgent action tendency is in effect trying to describe what all actions which *would* satisfy the tendency have in common.

The fact that actions are non-arbitrarily associated with urgent action tendencies is what allows us to maintain that an umotion can be identified with an action tendency, despite the fact that the tendency may both be inhibited and flexibly manifested in many different ways. The discussion of what exactly would satisfy a given urgent action tendency is to be made on a

case by case basis. Most commonly, an urgent action tendency will require physical actions for being satisfied. For example, sometimes the anger we experience towards someone who slighted us requires that we actually do something that will result in harm to them. In other cases, however, a tendency may be satisfied merely by mental actions. There are forms of anger which are perfectly well satisfied by merely *imagining* inflicting harm on somebody. Let us now turn to what has historically been the least explored of the three functional components of umotion, namely that of communication.

10.2.9. Communication

Under the view I am proposing, an umotion is a system for dealing with situations of urgency. These situations commonly emerge when there is a sudden change in the environment. Since other organisms represent one of the most important sources of sudden change in the environment, we must expect that the dynamic interaction between organisms will play a role in the activation and management of urgent action tendencies, and therefore in emotional episodes. The appreciation that umotions unfold dynamically through time is my main rationale for assimilating them to urgency *management* systems.

One of the key aspects of this management is related to *communication*. As I mentioned before, *communication* in umotion can be to the *self* or to *others*. Sometimes umotions deliver – at least to human beings - what Peter Goldie (2004) called “introspective knowledge”. By getting angry about being denied tenure, for example, one may realize how much one cares about an academic career, despite having claimed the contrary for a long time. This form of communication to the self is primarily achieved through feelings.

Often feelings will be feelings of bodily changes, but on occasion feelings will be feelings *of* an emotional appraisal and/or *of* an urgent action tendency, without involvement of any bodily changes (understood as autonomic changes). A feeling of fear, for example, can in principle comprise consciousness of appraising something as to be avoided, and consciousness of the behavioral tendency of avoidance, without any consciousness of increased heart beat and trembling, for the very good reason that these may not occur. This would be the case for example in cases of feelings of fear of global warming. In turn, consciousness of the behavioral tendency of avoidance may comprise consciousness of *preparing* for avoidance and consciousness of *acting*, for example through consciousness of one’s expressions (e.g. consciousness of the jaw

dropped open), of the physical action of, say, running and of the mental action of imagining what it would be like to be hurt.

Despite its importance, I leave the communication to the self in the background in what follows. I am more interested in forms of umotional communication to others, which are not much explored in emotion theory, with the exception of social constructionist theories of emotion and hybrid programs which take their cue from them (see chapter 6). To understand that umotions have a communicative dimension is to peel a further layer away from the picture of emotions as *happenings* which has loomed large in the history of the subject. This is because communicating is something one *does* by umoting. What communication adds to the picture sketched so far is a *negotiating* dimension to emotional phenomena, which is crucially related to how others respond to the unfolding of an umotion throughout *preparation* and *action*.

The idea that umotions have a negotiating dimension comes from ethology. In section 6.3.3, I discussed Hinde as a founding figure in the conceptualization of emotions as negotiating moves. Hinde (1985a, 1985b) suggested that emotional behaviors lie on a continuum between a purely expressive end and a purely negotiating end. This view was opposed to a research tradition Hinde associated with Darwin, according to which emotional behaviors have a one-to-one correspondence to the motivational states underlying them. Hinde argued instead that it makes evolutionary sense for organisms to probe the likely reactions of other organisms they interact with by sending out partially ambiguous behavioral signals and dynamically gathering information from the way such signals are responded to. By emotional behavior, Hinde understood mostly facial and postural expressions, but we need to expand the notion to all features of an urgent action tendency that can in principle communicate that the tendency is undergoing. Anything which characterizes the phases of *preparation* and *action* can in principle count as an signal, provided a recipient can use it to reliably infer that a certain urgent action tendency has been activated and is being managed.

Although emotion theorists have mainly focused on emotional expressions, expressions are not the only channel through which the activation of an urgent action tendency can be communicated. The activity of searching for a weapon in an expressionless fashion is certainly as informative of the occurrence of anger as the combination of a fixed stare, eyes widened, eyebrows lowered, bared teeth, and compressed lips. Generally, emotional communication

comprises several sources of signals working together. Facial, postural and vocal expressions are very important signals, but I think it is a mistake to focus on them exclusively when trying to understand the dynamics of emotional communication.

In any event, Hinde's basic point was that there is no one-to-one correspondence between "emotional states" and "emotional behaviors". This manner of speaking suggests that "emotional states" are one thing and "emotional behaviors" another, which raises the question of how we can identity one independently of the other (Hinde did not provide an answer to such question). In the context of UMS theory, "emotional states" are activations of a system for the management of an urgent action tendency, so the dichotomy between "emotional states" and "emotional behaviors" has no real grip.

At the same time, Hinde's insight on negotiation can still be captured by stating that there is no one-to-one correspondence between the activation of a given urgent action tendency, and what the emoter does to manifest it. As I argued in the section on emotional *action*, this is true not only in the sense that there are many actions which can be flexibly undertaken, but, more radically, that the tendency can be *inhibited* despite its urgency.[33] Hinde's insight boils down to the idea that the system which manages an urgent action tendency, namely umotion as I understand it, is sensitive to the relational goal affordances for its advantageous manifestation *as they are changed throughout the process of emotional communication*. Emotional communication, it must be emphasized, is not restricted to the preparatory phases of the management of an action tendency, but extends to what one communicates by acting upon detected affordances. Quite obviously, I receive an emotional signal when I observe the baring of teeth, the searching for a weapon, and the physical action of being shot at.

A possibility worth mentioning is that in some instances of umotion communication may even determine *what* relational goal is being pursued, not only *which means* to such goal will be chosen (if any). Some species of umotion could in principle start with the activation of an urgent action tendency whose relational goal is partially undetermined, and will be fully determined once responses to early emotional signals are received. For example, Hinde reports that birds often display threat expressions and subsequently flee (see my discussion in subsection 6.3.3). We can interpret such birds as being in a state of "emotional uncertainty", in which they manage an urgent action tendency which is hanging between the relational goals of *avoiding* and

[33] This is true only when umotion does not work a reflex

attacking. Reactions to threat displays will determine if the bird shifts to an urgent action tendency of the anger type or of the fear type.

An empirical question I won't pursue here is how often emotions begin with appraisals which only endow them with a partially undetermined relational goal. For now, I will consider this to be a special case of umotion, and assume that the standard case of umotion is that of the emergence of a fully formed urgent action tendency with a specific relational goal.

But what is communicated in the course of the management of an emotional episode? The quick answer is: Anything that can reliably be inferred from it. Most importantly, the activation of an urgency management system sends signals about its likely cause and its likely effects. The former include signals about what kind of emotional appraisal has taken place. For example, inferring that someone is managing an urgent attack tendency towards us communicates that they have appraised our conduct negatively, and that they most likely consider themselves to have been slighted by us. But realizing that someone is engaged in an urgent action tendency also tells us about what they are likely to do next, namely choose one of the various actions which fulfill the relational goal of damaging us.[34]

Receiving such signals is likely to affect our own behavior, for example by generating an umotion in us. This umotion will in turn have a communicative dimension, and send signals to the original emoter concerning our own evaluation of the situation and behavioral intent. This will generate a dynamic interaction familiar to all of us from our personal lives but very little studied from a scientific point of view (but see Parkinson et al. 2005 for a welcome exception). I conclude with a few examples of the kinds of signals which may be inferred from some possible urgent action tendencies:

[34] Not all of the signals sent by an umotion, it must be emphasized, are sent purposefully, namely as part of the proper functioning of the umotion (more on this shortly).

Urgent action tendency	*Communicative agenda*
Attack tendency	I have appraised your conduct as blameworthy, I am going to hurt you
Avoidance tendency	I have appraised my situation as dangerous, I am going to fight or flee
Reparation tendency	I have appraised my own conduct as blameworthy, I am going to do something to make amends to you
Appeasement tendency	I have appraised my conduct as falling short of a standard I endorse, I am not going to do it again

Figure 14: The communicative agenda of action tendencies

Let us now work with the account of umotions I have sketched, and tackle the last piece of our puzzle, which concerns the issue of emotional intentionality.

10.3. THE INTENTIONALITY OF UMOTIONS

The theory I articulated in sections 3.2.1 to 3.2.7 states that umotions are systems for the instantiation and management of an urgent action tendency, in brief urgency management systems. This theory, however, will not do as it is. What is missing from it is an account of portion (b) of the definition of umotion I provided in section 3.2:

An umotion E is a superordinate system, generally activated by an emotional appraisal, which:

(a) controls a cluster of organismic subsystems whose synchronized operation instantiates, and manages through time, an action tendency T_E with control precedence

(b) has a pragmatic object, which describes the conditions of pushmi-pullyu appropriateness PP_E for E

What we now have to understand is what a pragmatic object is, and what role it plays in the individuation of umotion-types. The notion of a pragmatic object is a development of Kenny's notion of formal object, which I described in chapter 4, and discussed again in 10.2.3. Here is a passage which can help us identify the problem we have to solve:

> If the emotions were internal impressions or behaviour patterns there would be no logical restrictions on the type of object which each emotion could have…In fact, each emotion is appropriate-logically, and not just morally appropriate-only to certain restricted [formal] objects" (Kenny 1963, 192)

Kenny's point was that a good theory of emotions must account for the fact that, in order to be appropriate, any given emotion must be about a certain restricted formal object. As I said before, by qualifying this notion of appropriateness as *logical*, Kenny meant to emphasize that he was talking about a sort of appropriateness which an emotion acquires just by virtue of being what it is. The basic idea is that, say, fear is not only *inappropriate* in the absence of danger, but *it is the sort of thing* which is inappropriate in the absence of danger. This condition of appropriateness contributes to identifying what fear is. Philosophers have a number of ways to designate entities with constitutive conditions of appropriateness of the kind emotions appear to have. They may say for example that emotions are *intentional* or *contentful* or that they have *aboutness* or the *capacity to represent*. I will understand all such notions as being equivalent ways of referring to the fact that emotions appear to have constitutive *norms of appropriateness*. The fact that emotions can *misrepresent*, namely can violate their intentionality-constitutive norms of appropriateness, is the central phenomenon I want to explain.

The idea that emotions have formal objects has always been the biggest albatross around the neck of both feeling theory and behaviorism. It is not hard to see why a theory assimilating emotions to either *mere* "internal impressions" or *mere* "behavior patterns" would have a hard time explaining why they have formal objects. If fear were *just* a feeling of bodily changes, then it would be mysterious why having such feeling in the absence of danger ought to be considered inappropriate.

A version of this problem would also afflict the UMS theory, if all it said was that fear is the instantiation and management of an *urgent tendency to avoid a certain object appraised as dangerous*. The question in such case would be: Why on earth would the *absence of danger* make it inappropriate, if it is only a tendency to avoid an object appraised as dangerous? The intuitive answer to this question strikes me as the right one, namely that the absence of danger makes fear inappropriate because fear has the *proper function* of being elicited by danger. But what exactly is the proper function of an umotion? And what sort of relation does the function of an umotion have to its intentionality? The following section is meant to provide a general framework for answering these sorts of questions.

10.3.1. Millikan's theory of intentionality: a sketch

One of the problems with developing a theory of *emotional intentionality* is that the nature of *intentionality* is independently problematic. The theorist is faced with an impressive variety of alternative accounts of intentionality, all endowed with some degree of plausibility, but fraught with their own domain of substantive problems. In this dissertation, I won't be able to explore alternative theories of intentionality. My strategy is to borrow pretty much whole what I take to be one of the most promising theories of intentionality currently available, namely Ruth Millikan's (1984, 1993, 2000, 2004) teleosemantics. I take no position on the controversies surrounding Millikan's theory as a general theory of intentionality. I will work with a "bare bones" version of her theory, introducing just enough detail to make use of it with respect to emotions, but without worrying about the many layers of philosophical sophistication intended for domains of application other than emotions (e.g. the emergence of language).

My objective is to show that Millikan's theory offers us a promising framework for making sense of emotional intentionality. To illustrate Millikan's theory, I rely on Millikan (2004), in which the theory presented in Millikan (1984, 1993) is complemented with a new theory of semantic information. This theory is offered in the context of Millikan's (2004) articulation of "a truce" with the position that intentional representations are purposefully produced natural signs that carry natural information (75-76). I will come back to the terms of this truce shortly, but let us begin by asking: What are natural signs? Millikan (2004) characterizes them as "vehicles that bear natural information", and offers a new theory of what is required for a vehicle to carry natural information. In the past twenty-five years, the standard account of natural information

has been Dretske's (1981). In broad outline, Dretske's theory states that X carries the information that Y insofar as (a) the conditional probability of Y given X is 1, (b) there is a law of nature or of logic underwriting the perfect correlation between X and Y.

The inspiring thought of Millikan's (2004) new theory of "locally recurrent natural information" is instead that "[n]early all of the kinds of information needed by us, and by all other organisms as well, for securing what we need in an inclement world, is information that cannot possibly be acquired without leaning on certain merely statistical frequencies" (32-33). The task for a theory of natural information becomes that of characterizing the properties of the statistical frequencies which allow for the transmission of information in the real, inclement world actual organisms inhabit.

The essential difference with Dretske's theory is that, instead of requiring information transmitting correlations to be perfect and nomically underwritten, Millikan (2004) requires them to be *non-accidental* within an appropriate reference class. But what reference class is appropriate for a theory of natural information? As Millikan notices, if there were no limitations on what counts as an appropriate reference class, we could always find a gerrymandered reference class such that, relative to it, a certain natural sign bears natural information about anything it correlates with it within such gerrymandered class.[35] This notion of information, however, would fail to tackle what Millikan (2004) takes to be the central task when developing a theory of natural information, namely explaining "why [an organism] might be able to use the recurrent sign as an indicator of its signified with some success" (38-39). To do so, Millikan suggests that we must consider *natural reference classes*, where a "natural reference class for a sign --the natural domain within which certain A's are "locally recurrent signs" of certain B's-- is a domain within which the correlation of As with Bs extends from one part of the domain to other parts for a reason" (40). Such reason will in turn be what explains why organisms learn about the signified from encountering the sign that carries information about it in the actual environments which they inhabit. As Millikan emphasizes, a correlation (defined relative to a natural reference class) does not need to be nomically underwritten or even strong for an organism to be able to use it to collect useful information. Her approach can consequently accommodate the fact that organisms can learn reliably from correlations with very different degrees of strength, as long as there is some reason why they hold in a natural reference class.

[35] I owe the clarification of this point to Ruth Millikan's comments on an earlier draft

For example, assume that there is a strong correlation between a certain type of sound X and a certain type of predator Y in a certain ecological niche N (e.g. X=sound of poisonous snake, Y=poisonous snake, N=area of Borneo forest). The correlation is not going to be perfect, and there certainly won't be a law of nature in N which makes the presence of Y nomically required given X. For example, it is physically possible and ecologically expectable that some organism other than Y may imitate sound X for strategic purposes (e.g. to avoid being attacked by predators), although rarely enough to preserve a strong correlation between X and Y in N.

Dretske's theory only allows us to conclude that this is *not* a case in which X carries the information that Y is present in N. Nevertheless, X obviously carries *some* information about Y's presence, namely that it is very likely. Millikan's (2004) theory has no difficulty accommodating this case. For X to carry (locally recurrent) natural information about Y, it is only required that the correlation between X and Y is not accidental within a natural reference class, in this case the ecological niche N. Such conditions would be fulfilled in case X was produced by snake Y most of the time in N. The correlation between X and Y would hold for a reason – X is caused by Y in N –, but it would nevertheless not be perfect, and would only hold locally in niche N. This notion of information strikes me as the one we need to develop a promising teleosemantic theory of emotional representation. From now on, when I speak of *information*, I will understand it as information in Millikan's (2004) sense.[36]

Let us now go back to the use Millikan makes of her new theory of information. As I mentioned before, the theory is formulated to make a "truce" with standard versions of informational semantics (e.g. Dretske 1986). The truce boils down to adopting a working account of intentional representations according to which *representations* (a.k.a. *intentional signs*) are "produced by systems designed to make natural signs for use by cooperating interpreting systems" (73). Under this view, representations emerge when some producer and some consumer (possibly parts of the same organism) are *designed to cooperate in the production and consumption of natural signs*.

Millikan does not fully endorse this account, however, but only uses is as a first approximation, useful for exploring the relation her theory of intentionality bears to

[36] I have argued elsewhere that thinking of information as a graded commodity amounts to capitalizing on one of the central lessons of Shannon's theory of information, and I have tried to offer a new measure of semantic information which takes that lesson to heart (Scarantino 2005). The theory of information I offered in Scarantino (2005) is broadly compatible with Millikan's own, but tries to go beyond it in the attempt to quantify how much information statistical correlations carry.

informational semantics.[37] The reason why intentional signs are not to be assimilated to natural signs purposefully produced is that there are some instances of intentional representations that fail to be natural signs purposefully produced. Millikan's (2004) biggest worry concerns intentional representations in which the correspondence between the sign and the signified does not hold for a reason, thereby not involving the transmission of natural information. Under such circumstances, the "correspondence of the sign to a real affair may be brought about by accident [and] the sign is not a natural sign of the condition that happens to satisfy it". Yet, an intentional sign mediates the cooperation between a producer and a consumer, and represents by virtue of an isomorphism between the sign and the signified (Millikan 1984).

Although not every intentional sign is also a natural sign, Millikan acknowledges that a great many intentional signs are, and that the production of natural signs may in fact be the "normal means" by which the producer generates true intentional signs. I will leave these caveats in the background, and work with the approximation that intentional signs are purposefully produced natural signs that stand midway between a producer and a consumer.

Millikan's central teleosemantic hypothesis is that the cooperation between producers and consumers of some natural signs exists because it had some beneficial effects on the producer and the consumer in a set of past circumstances by virtue of which the sign production-sign consumption cooperation was selected for, relative to some selection mechanism. Millikan designates such past circumstances of selection with the label "Normal". As I understand Millikan's (2004) theory, there are no narrow limits to the type of selection mechanism at work in Normal circumstances. As she puts it to mention two prominent cases, some mechanisms of sign production "have been selected for by natural selection during evolutionary history. Others are selected for or tuned for their jobs through processes of learning".

What is required for something to count as a beneficial effect is that it must explain why a certain producer started generating natural signs and a certain consumer started using them, thereby establishing a partnership which was to the advantage of both in Normal circumstances.[38] This allows for beneficial effects to comprise fitness benefits, the avoidance of pain in a conditioning experiment, the fulfillment of social functions, and so on. Complexities

[37] I owe the clarification of this points to comments by Ruth Millikan on an earlier version of the chapter

[38] Millikan (2004) seems to hold an even weaker notion of *benefit*, according to which what counts as a benefit is anything which explains why the sign production-sign consumption cooperation was *not selected against*. I will not consider this wrinkle of her theory in what follows.

aside, the beneficial effect responsible for selection can be designed as the effect the sign production mechanism has when it fulfills its *proper function* (see Millikan 1984, 1993 for a much more nuanced account). According to Millikan, it makes sense to speak of intentional representations when the production of natural signs has acquired a proper function.

I endorse this approach to intentionality under a couple of presuppositions, namely that no assumptions are made about the time in which circumstances were Normal, nor about the mechanism of selection at work in such circumstances. As far as I am concerned, Normal circumstances can be arbitrarily close in time to the current circumstances, as long as it makes sense to speak of a selection process as having taken place. Under this liberal understanding of proper functions, the acquisition of the ability to represent and the emergence of failures to represent need not be separated by evolutionary ages.

Moreover, I assume that the same mechanism can have different proper functions relative to different sets of Normal circumstances. For example, a mechanism may have acquired a certain proper function through natural selection relative to Normal circumstances very far in the past, and have acquired another proper function through learning relative to a more recent set of Normal circumstances. A full articulation of these qualifications would take me too far, but I take them to be broadly compatible with Millikan (2004).

What matters mostly to my project is a distinction Millikan makes between three kinds of representations or intentional signs:

> [D]escriptive intentional signs…are designed to stand in for world affairs, typically affairs outside the organism, and to vary according to these world affairs (2004, 80).
>
> Directive signs guide the consumer in the production of world affairs that vary according to how the signs themselves vary. They are blueprints for what is to be constructed or brought about (2004, 80).
>
> [Pushmi-pullyu] signs…are signs that are undifferentiated between presenting facts and directing activities appropriate to those facts. They represent facts and give directions or represent goals, both at once (2004, 157).

The distinction is grounded on differences between what natural signs are produced *for* in the context of the producer-consumer partnership. Descriptive signs, Millikan argues, ought to

"stand for world affairs". To use a turn of phrase coined by Austin (1962) and borrowed by Searle (1983) to account for differences between classes of propositional attitudes, descriptive signs have a *mind-to-world* direction of fit, in the sense that their constitutive aim is to have a content which *fits what the world is like*. There are many accounts available of the "fitting relation", but the point is that the proper function of a descriptive sign is to offer the consumer a sign whose beneficial use in Normal conditions is contingent upon standing for world affairs. The responsibility for the fit between the world and the sign, we may also say, is on the shoulders of the producer of descriptive signs (Searle 1983).

Imperative signs, on the other hand, ought to "guide the consumer in the production of world affairs". Their direction of fit is *world-to-mind*, in the sense that their constitutive aim is to have a content which offers a blueprint for *what the world is to be made like*. In other words, the proper function of an imperative sign is to offer the consumer a sign whose beneficial use in Normal conditions is contingent upon motivating the consumer to change to world in such a way that the fitting relation between the sign and world affairs is brought about. The responsibility for the fit between the world and the sign, to carry on with the metaphor, is on the shoulders of the consumer in the case of imperative signs.

Pushmi-pullyu (PP) signs, finally, "are undifferentiated between presenting facts and directing activities appropriate to those facts". They have what we may call a *dual direction of fit*. The constitutive aim of such representations is to offer the consumer a sign whose beneficial use in Normal conditions is contingent upon *being generated by the producer in circumstances in which it motivates the consumer to act as specified by the sign's content*. Such content must be understood under the presupposition that there is no separation of the purpose of *varying so as to fit the world*, as descriptive signs do, and the purpose of *guiding behavior according to how they vary*, as imperative signs do. PP signs unify descriptive and imperative dimensions, roughly in the sense that their purpose is to *guide behavior according to how the world varies*. In other words, there is no differential allocation of the responsibility for fitting between producer and consumer.

Under this account of the distinction between signs, descriptive and imperative signs must work together to generate behaviors, each bringing to the table of their cooperative endeavor its own distinct content. For example, to lead someone to go to Paris, the belief that Paris is the capital of France and the desire to go to the capital of France must work together. On the other

hand, PP signs can generate behavior on their own, without requiring any collaboration between a *descriptive content* which "presents facts" and an *imperative content* which "gives directions".

The distinction between the *content* of descriptive and imperative signs and the *content* of pushmi-pullyu signs can be appreciated when we try to get a handle on their conditions of appropriateness. Whereas descriptive signs are signs whose proper function is *to have a true content*, and imperative signs are signs whose proper function is to have a *satisfied content*, i.e. a *made-true* content, pushmi-pullyu representations are signs whose proper function is *to have a content which guides behavior appropriately in the circumstances in which it is produced*. To emphasize this important distinction, I call the content of descriptive and imperative signs *semantic*, and the content of pushmi-pullyu signs *pragmatic*.

Notice that this account does not presuppose that the mechanism producing descriptive, imperative and pushmi-pullyu representations will only produce representations which manage to achieve what they ought to according to their proper functions. As far as Millikan's theory is concerned, representation failure can be widespread for all kinds of representations. I will call a representation which fails to fulfill its proper function *false* when the representation is descriptive, *unsatisfied* when it is imperative, and *pragmatically inappropriate* or *pushmi-pullyu inappropriate* when it is a pushmi-pullyu representation.

Failure for a PP representation amounts to being produced in circumstances in which it does not generate the kind of benefit that established the producer-consumer partnership in Normal circumstances relative to a certain selection mechanism (e.g. natural selection, learning, etc.). This can happen either because the PP representation is produced in circumstances other than the ones by virtue of which it was selected for in Normal circumstances, or because it guides behavior in ways other than the ones by virtue of which it was selected for in Normal circumstances, or for both reasons at the same time. This is admittedly just a sketch of a very complicated theory, but it will hopefully do for my purposes.

10.3.2. Umotions as pushmi-pullyu representations

There are currently two main accounts of emotional intentionality in contemporary emotion theory, and they both seem to me inadequate. One is provided by cognitivists, who argue that emotions have intentionality in the same sense in which judgments have intentionality. For example, the intentionality of fear is characterized as the intentionality of the judgment that danger

is present. In the absence of danger, such judgment is false, and fear consequently inappropriate. This popular view either misrepresents or fails to illuminate the phenomenon of emotional intentionality (see chapter 7 for a more extended discussion).

It misrepresents it if we understand judgment as the paradigmatic manifestation of linguistic/conceptual abilities, as the expression of a central rather than modular information processing system, and as a cognitive attitude with a mind-to-world direction of fit. This is because emotions can occur in creatures without language, they can be elicited by modular forms of appraisal, and they always have a motivational dimension.

On the other hand, it fails to illuminate it if we apply to the notion of judgment what in chapter 7 I called the *Placeholder Strategy*, according to which the notion of judgment is expanded to accommodate all properties emotions may be taken to have. Under such interpretation, to say that an emotion *E is the judgment that p* is simply to say that p is a description of E's conditions of appropriateness. What remains unaccounted for is the relation between the emotion and such conditions of appropriateness, which is what a theory of emotional intentionality should explain.

A promising alternative to the cognitivist understanding of intentionality has recently been proposed by Jesse Prinz (2004a, 2004b). According to Prinz's Neo-Jamesian theory, emotions are perceptions of bodily changes which represent by virtue of their function. More precisely, Prinz has argued that the *vehicle* of emotional representation are bodily changes, and what grounds the *representation relation* between emotions and what they are about is that such bodily changes have the function of being caused by specific eliciting circumstances they consequently come to represent. If Prinz is correct, fear is not a judgment that danger is at hand, but rather a (conscious or unconscious) perception of fear-typical bodily changes (e.g. quickened heart beats, shallow breathing) whose function is being caused by danger.

I have three main problems with Prinz's account of the intentionality of emotions. The first is with the assumption that bodily changes are necessarily the vehicle of emotional representation. I have argued that there are many cases of emotions which do not involve such changes (see chapter 7).

The second problem concerns the details of Prinz's theory of mental representation, more specifically the theory of information it presupposes. Prinz relies on Dretske's (1981, 1986, 1988) theory of mental representation, according to which mental states represent what they have

the function of carrying information about. The problem is that the *information carrying relation* between two events is assumed to amount to a nomically underwritten perfect correlation between them. But this is not a viable account of how emotions carry information. The sense in which, say, fear carries information about dangers is not that there is a law of nature according to which whenever fear is instantiated danger is instantiated. Rather, the idea is that there is an imperfect but not accidental correlation between fear and danger by virtue of which the former carries information about the latter. By relying on Dretske's theory, Prinz inherits all the limitations of Dretske's (1981) account of information.

The third and most significant problem is that Prinz's theory characterizes the emotions as *descriptive representations*, leaving their key directive side in the background. According to the theory he proposes, fear represents danger because it has the function of being correlated with danger. But the function of fear is not merely to *track* the presence of danger, as the belief that danger is present aims to do, but rather to *direct behaviors appropriately when danger is present*. Under the view proposed by Prinz, the intentionality of emotions is ultimately grounded in the idea that emotions are *bodily changes with the function of correlating with specific states of affairs*. But it is unclear what kind of function this sort of correlation could serve. What would be the advantages for an organism of undergoing bodily changes *as such*?

The three limitations I have singled out in Prinz's account can all be resolved if we assume that the vehicles of emotional representation are umotions, and if we apply to them Millikan's (2004) theory of pushmi-pullyu representations.

My central thesis is that umotions are urgency management systems endowed with a special kind of pushmi-pullyu (PP) intentionality. This is to say that urgency management systems are intentional representations with a *pragmatic* rather than *semantic* content. Under the premises of Millikan's theory of intentionality, to say that umotions are intentional representations is to assume that they stand midway between a producer and a consumer which are designed to cooperate. This in turn presupposes that there were Normal conditions such that, relative to some mechanism of selection, a producer and a consumer of any umotion E were selected for cooperating in the production and use of E. The beneficial effects explaining why selection took place in such Normal circumstances account for the *proper function* of the urgency management system that is E (relative to a given selection mechanism).

Now, what are the designed producers and consumers of umotions? The designed producer of umotion is the mechanism I called *appraisal*. Although umotions may occasionally be produced by mechanisms other than appraisal, it is quite clear that these are not the forms of elicitation that led to the establishment of the cooperation between producer and consumer. What alternative forms of elicitation such as, say, direct brain stimulation, facial feedback and chemical induction lack, as I mentioned before, is the monitoring capacity of appraisal, which is key to eliciting umotion E in circumstances in which its characteristic urgent action tendency T_E is beneficial. The designed consumer of emotion is instead the cluster of subsystems umotion controls, and, more distally, the emoter who hosts them. I organized such subsystems into functional components, so we may say that the subsystems *consume* emotions for purposes of preparation, action and communication. Since both emotional appraisal and the subsystems governed by umotion are part of the emoter, we can refer to umotions as *inner intentional representations*.[39]

For the purposes of my analysis, the proper function of an umotion can be due to all sorts of advantageous effects, ranging from benefits in terms of evolutionary fitness to benefits in terms of learning in a conditioning experiment. An account of the umotions' intentionality cannot be given in general, because there is no reason to suppose that a unique mechanism of selection explains why every urgency management system available to an organism was selected for. Just to give an example, "fear" and "disgust" are two excellent candidates for being umotions selected by natural selection. As I argued in chapter 5, the evidence on the neurobiology and facial expressions associated with fear and disgust points to their being biological adaptations. Some other umotions, however, are most likely the result of a process of cultural selection. For example, guilt may conceivably have been selected for because of the benefits it bestows to agents by making them capable of abiding by a system of mutually beneficial social norms.

The task of this section, however, is not to investigate the proper function of any single umotion in depth, but rather to sketch a general theory of emotional intentionality. What makes the idea that umotions have proper functions at least prima facie plausible is that urgency management systems have properties which are likely to have been (and still be) beneficial in

[39] This is compatible with holding that an umotion may also qualify as a *non-inner representation* with respect to the relation between the emoter and another interactant. For example, emotional expressions are used also by organisms other than the emoter. Also, holding that umotions are inner representations is compatible with the possibility that the benefits to the emoter that explain why the cooperation between producer and consumer was established involve other organisms, as for example in kin selection.

many classes of circumstances. This is what makes the exploration of what such circumstances are for each umotion E worth pursuing, a task I leave for a future time.

The potential advantages of umotions appear to be related to the peculiar action control structure an urgency management system embodies, paradigmatically characterized by a combination of speed and flexibility. This sort of combination is exactly what organisms need to deal with complex environments. Organisms only capable of the speed of reflexes would in such circumstances fare very poorly, because they would be unable to adapt their responses to a rapidly changing environment. On the other hand, organisms only capable of the flexibility of practical reasoning would also find themselves in trouble, because performing cost-benefit analysis in rapidly changing circumstances leads in many circumstances to failing to act quickly enough.

Although speed and flexibility lead us to naturally conceptualize the advantages of umoting in terms of the functional components I called *preparation* and *action*, we must not forget the potential benefits of *communication* when we explore the proper function of emotions. The UMS theory assumes that the *proper function* of a given umotion E depends upon the combination of benefits accruing (in Normal circumstances) from the *integrated combination* of emotional *preparation*, emotional *action* and emotional *communication*. For example, part of the reason why fear was selected for will be that there are advantages related to communicating to one's kin the presence of danger. Similarly, part of the reason why anger was selected will include that angry emoters communicate hostile behavioral intentions to others, often achieving submission without actually having to engage in potentially costly fights.

The main novelty of the account of emotional intentionality I propose, however, is not that emotions have proper functions. Rather, the main novelty is that umotions are characterized as *pushmi-pullyu representations*, whose proper function is to present facts and direct behaviors *at the same time*.[40] Umotions are clearly not the only forms of pushmi-pullyu representations. For example, Millikan describes bee dances as pushmi-pullyu representations, which at the same time tell spectator bees where the nectar is and direct action towards getting it. One may think that the relevant difference is that umotions are inner representations, in which producer and consumer exist within the same organism, whereas bee dances are non-inner representations, in which producer and consumer are distinct organisms. But this won't do, because there are

[40] I owe the development of this idea to discussions with Ruth Millikan

innumerable inner PP representation other than umotions. For example, Millikan characterizes "intentions to act" as PP representations, because they serve "at once to direct action and to describe one's future so that one can plan around it", and "intentions" are undoubtedly inner.

What distinguishes umotions from other kinds of inner pushmi-pullyu representations, I argue, is that they have a special way of *pushing* and a special way of *pulling*. Concerning the pushmi or directive side, umotions are designed to direct behavior *in the style of urgency management systems*, namely with control precedence. *Intentions* as such do not have control precedence, in the sense that they are states of readiness to act, possibly also of the "intention in action" variety (Searle 1983), but not endowed with the sort of multi-layered "clamoring for attention and execution" which I singled out as the mark of the emotional.

Concerning the pullyu or descriptive side, umotions are designed to be produced by an *appraisal*. This is what allows us to distinguish them from PP drives such as hunger and thirst, which are inner PP representations but not umotions. The designed mechanism of elicitation for hunger and thirst, in fact, is a self-regulating biological mechanism that operates in rhythmic cycles of detection and stabilization of physiological imbalances. Emotional appraisal, on the contrary, monitors sudden changes in the environment, and does not rely on a homeostatic internal system. The *pullyu side* of umotion, in other words, aims to track events which demand prioritized goal pursuit but whose emergence is not cyclical.

To say that umotions are PP representation, I emphasize, is not to imply that they will always manage to achieve what they ought to according to their proper function. What is implied is only that there were Normal circumstances in the past such that the mechanism producing and consuming umotion E produced tokens of it which guided action in the circumstances in which they were produced, and had beneficial effects which explain why the production-consumption partnership was selected for with respect to E (by natural selection, by learning, etc.).

What identifies umotion types, in conclusion, is a conjunction of a certain urgent action tendency T and of a certain set of conditions of pushmi-pullyu appropriateness PP, which in effect characterize the circumstances in which the urgent action tendency T fulfills its proper function. The UMS theory offers a novel identity condition for emotions, and it explains the sense in which urgency management systems can acquire a normative dimension. When we know the conditions of PP appropriateness, we know what an umotion E is supposed to do, and we can speak of *misrepresentation* when it fails to do it (possibly most of the time).

I call the description of such conditions of appropriateness the *pragmatic object* of an umotion E. By speaking of pragmatic objects, I mean to signal that I am referring to conditions that are constitutively associated to a given umotion and can contribute to type-identifying it. This is Kenny's (1963) idea of formal objects, with a couple of twists. The first is that the constitutive relation between an emotion and its object is grounded in a history of selection. To discover the pragmatic object of an emotion, what we need to do is to understand its proper function, which tells us under what conditions a given emotion E does what it ought to. The second is that the object is to be understood in terms of pragmatic rather than semantic conditions of appropriateness.

Since PP representations do not distinguish between presenting fact and directing activities, the idea of pragmatic objects demands a new understanding of emotional content. Emotions which do not fulfill their constitutive conditions of appropriateness are generally described as being *false* in emotion theory (e.g. De Sousa 1987), but this manner of speaking has two shortcomings. On the one hand, it wrongly suggests that the descriptive content of an umotion is represented in the same way in which the content of a belief is represented. This interpretation is almost inevitable when emotions are identified with judgments, as in the cognitivist theory of emotions. On the other hand, to speak of representational failure for emotions exclusively in terms of falsity neglects the directive side of umotion, as if an umotion could tell what is the case without at the same time telling what to do about it.

According to the UMS theory, the proper function of umoting is instantiating the match between conditions of elicitation and behavioral guidance that allowed the production-consumption cooperation of umotions to be selected for (relative to some mechanism of selection). Since I allow for the possibility that the same mechanism may have proper functions with respect to different selection mechanisms (e.g. natural selection and cultural selection), it is possible that an umotion may acquire more than one kind of pragmatic content at different times, a complication I will not further explore.

The intuitive relation between emotions and their formal objects, described by Kenny (1963) as being what a theory of emotions as behavioral predispositions could never explain, is now perfectly explainable. Under the UMS theory, the reason why fear is inappropriate in the absence of danger is that danger is a description of the conditions of PP appropriateness under which the superordinate system producing urgent avoidance tendencies was selected for. Let me

emphasize again that these may or may not be the circumstances in which an emoter is currently disposed to produce fear. The point is that, even if an emoter is disposed to engage in urgent avoidance tendencies with respect to non-dangerous circumstances, the system generating tokens of such tendency was selected for doing so when danger was present. This is what allows us to conclude that engaging in urgent avoidance tendencies in the absence of danger instantiates a form of *malfunctioning* for the fear system.

In a nutshell, the pragmatic object of an emotion E, what emotion theorists have tried to capture with notions such as *formal object* (Kenny 1963) or *core relational theme* (Lazarus 1991), is a description of the conditions under which E has the proper function of being elicited.

Characterizing pragmatic objects linguistically is difficult, because their content is both descriptive and directive but does not result from a combination of a purely descriptive representation (e.g. a belief) and of a purely directive representation (e.g. a desire). The proper function of umotions is neither to independently *vary so as to fit the world* nor to independently *guide behavior according to how it varies*, but rather to *guide behavior according to how the world varies*. There are two ways of failing to achieve this objective. On the pullyu side, an umotion may be activated in the wrong circumstances. For example, "fear" may be activated by events which are not dangerous. On the pushme side, "fear" may direct action in the wrong way. Even though activated by an actual danger, "fear" may bring about behaviors which are not an adequate response to it (e.g. excessive trembling).

At the heart of the idea of pragmatic content as I understand it is the idea that these two forms of failure are equivalent as far as the pushmi-pullyu representativeness of an umotion is concerned. Failing to PP represent is either failing to be produced in the state of affairs in which consumers were historically benefited or failing to guide consumers in the historically beneficial production of world affairs or both. Differently from descriptive and imperative representations, the responsibility for the fit between intentional signs and world affairs is on the shoulders of both producers and consumers of PP representations.

10.3.3. Are all emotions umotions?

I do not believe that all things we call emotions in ordinary language are umotions, nor that all things we call fear, disgust or anger in ordinary language are umotions. Far from being a shortcoming of my theory, I take this aspect to be one of its major strengths. I argued in chapter

8 that the domain of vernacular emotions is highly heterogeneous and vague. An emotion theorist can either try to capture the family resemblance characteristic of folk emotion categories (Folk Emotion Project), or try to develop an explication of them (Explicating Emotion Project). As anticipated in the introduction, I offer the UMS theory of emotions in the spirit of an explication. Umotion does not make explicit what "emotion" means in ordinary language, but rather what is a "good thing to mean" by it when we engage in the intellectual pursuits of scientific psychology.

In other words, my aim has been to transform the folk category of "emotion", and the folk categories of specific emotions such as "fear", "anger" and "disgust", to make them suitable for scientific psychology while maintaining enough similarity with the folk categories to count as explicating them. Minimally, this task requires eliminating some of the vagueness characterizing folk emotion categories, and offer fairly exact ground rules for their use. The ultimate objective of the UMS theory, however, is more ambitious: it is the substitution of folk emotion categories with scientific categories *fruitful* for the purposes of scientific psychology. This substitution is not meant to eliminate "folk emotion" or "folk fear" from ordinary language, a task which is both meaningless and unachievable, but rather to eliminate such categories from the language of working scientists.

I will not be able to demonstrate that umotion is a fruitful category, a task which can only be achieved once the theoretical construct of umotion begins being used by psychologists, biologists and neuroscientists in their own intellectual pursuits. On the other hand, I believe I have clearly reduced the vagueness of the folk category "emotion", and articulated a theoretical concept which at the very least carries some promise of being fruitful (for the purposes of scientific psychology). Let me just give one example of this potential fruitfulness, comparing the UMS theory with cognitivism.

Two of the most valuable avenues of research in contemporary emotion theory involve comparative animal studies and the study of ontogenetic emotional development. One of the central puzzles faced by these research programs is that it is unclear in what sense adult humans, infants and animals can all experience the same emotions, given their differences in cognitive sophistication. In the absence of an account of emotions we can use to make sense of the continuity of emotional phenomena across species and stage of ontogenetic development, these research programs lie on shaky foundations.

The cognitivist account of emotions as judgments fails to offer a theoretical construct usable in these scientific studies. It is unclear how pre-linguistic and non-linguistic creatures could possibly issue *judgments* without mastering a language. The standard cognitivist response of trivializing the notion of judgment so as to make it available *by fiat* to infants and animals (see 7.2.1) fails to shed light on what it is that adults, infants and animals share when it comes to emotional phenomena. The construct of umotions offers instead a viable answer, even though certainly not the only one (see below). Adults, infants and animals can all emote in the sense that they can all engage in urgent action tendencies generated by a superordinate system endowed with a proper function.

Cognitive differences between them will affect what events are appraised in a way that brings about an umotion. On the other hand, there is nothing mysterious about the idea that creatures of different levels of cognitive sophistication may be able to engage in, say, the urgent avoidance tendency of "fear". I have so far used terms such as "fear", "anger" and "disgust" to illustrate examples of umotions, but it is now time to make explicit that not all items comprised in such subordinate folk categories qualify as umotions. From now on, I will speak of "ufear", "uanger" and "udisgust" when I intend to refer to the *explicata* of subordinate folk emotion categories proposed by the UMS theory.

In section 9.2.1, I reformulated the Carnapian account of explication, arguing that the reduction of vagueness and/or the acquisition of fruitfulness (relative to some theoretical objectives) are the two fundamental payoffs attached to explication. I also said that no explication can be good unless it achieves similarity in use with the explicandum. The question is now: Is it the case that "in most cases in which [emotion] has so far been used, [umotion] can be used"? In other words, is "umotion" interchangeable with "emotion" (salva veritate) in enough linguistic contexts to achieve "similarity in use" with it? It seems to me that umotion passes the similarity test with flying colors: in very many linguistic contexts in which the term "emotion" is ordinarily used what is being referred to is precisely an umotion. At the same time, it is also clear that there are many cases of ordinary emotions which are not captured by the UMS theory.

In the chart below, I report a few examples of paradigmatic umotions. Umotions are identified in the style of the UMS theory, namely by means of a conjunction of an urgent action tendency and its conditions of pushmi-pullyu appropriateness. This criterion of identification is different from the ones used by cognitivists, according to which emotions are type-identified by

judgments, and by Neo-Jamesianism, according to which emotions are type-identified by perceptions of bodily changes. In the last column, I give some examples of the sorts of benefits which may explain why a given umotion was selected for (relative to some mechanism of selection):

Umotion	*Urgency management system for the management of an urgent action tendency of…*	*…in circumstances of PP appropriateness characterized by….*	*…with a proper function connected to…*
Uanger	… attack/obstacle removal	…slight/something resisting our getting through	Preparing for obstacle removal and/or Executing actions of obstacle removal flexibly and quickly and/or Communicating signals of negative other-evaluation and aggressive behavioral intent
Ufear	… avoidance	…danger	Preparing for evasive action and/or Executing actions of avoidance flexibly and quickly and/or Communicating signals of danger and need for help
Uguilt	…reparation	…violation of important norms of conduct	Preparing for reparation and/or Executing actions of reparation flexibly and quickly and/or Communicating signals of negative self-evaluation and need for forvigance
Uembarassement	…appeasement	…violation of relatively unimportant norms of conduct	Preparing for appeasement and/or Executing actions of appeasement quickly and flexibly and/or Communicating signals of negative self-evaluation and need for avoiding attention

Figure 15: Some examples of umotions

I intend my descriptions of the tendency T_E and of the conditions of appropriateness PP_E of any given E to be only tentative. A thorough analysis would demand a complex study of the proper function of each umotion E. This is because we cannot establish what the pragmatic object of, say, anger is without having first understood what the proper function of anger is. This in turn requires asking under what sorts of circumstances the urgent action tendency associated with anger had benefits for the emoter which explain why it was selected for. The chart offers a few tentative descriptions of urgent action tendencies, conditions of pragmatic appropriateness and benefits that may in principles explain why a given umotion acquired a proper function.

For example, “uanger” is identified with an urgent tendency of attack or obstacle removal, pragmatically appropriate in case a slight has been committed or something resists our getting through. The benefits associated with “uanger” are related to the three functional components of preparation, action and communication that are characteristic of it. For example, it is intuitively persuasive that the system producing tokens of anger was selected for because it efficiently prepared for vigorous hostile action, because it allowed for a quick yet flexible action with respect to an obstacle, and because it sent signals of negative appraisal and hostile behavioral intent that helped bringing about submission.

Similar stories can be told with respect to the possible benefits of “ufear”, an urgent action tendency of avoidance pragmatically appropriate in the presence of danger, “uguilt”, an urgent action tendency of reparation pragmatically appropriate when an internalized and important norm has been violated, and “uembarassment”, an urgent action tendency of appeasement pragmatically appropriate when one has made a faux pas involving the violation of relatively unimportant norms of conduct.

It is clear from the simple mastery of the English language that very often when we speak of anger, fear, guilt, embarrassment and many other emotions, what we are referring to are precisely the sorts of urgency management systems summarized in the chart above. A paradigmatic case of anger, say a state of intense anger generated by being denied tenure and mocked for it by a colleague who announced us the bad news with a smile, is an urgent attack tendency which comprises the preparation of the body for action, the execution of mental, physical and expressive hostile actions, and the communication of the appraisal of having been unjustly treated and of currently being belligerently disposed. The attack tendency can be

flexibly manifested, depending on the circumstances, and it can also be inhibited. It would be pragmatically inappropriate if no slight had been committed, or more generally if no obstacle had been put by anyone or anything on the path of the emoter.

At the same time, there appear to be cases of anger which are not "uanger". For example, the sort of mild and self-righteous anger one may experience in the morning while sipping a coffee and reading about politicians' squabbles need not consist of any urgent attack tendency. Similarly, there will be cases of fear which are not "ufears". For example, one may speak of unconscious fear of failing in life as a way to explain a complex pattern of avoidance behaviors displayed in diverse occasions throughout a number of years. It is unclear in what sense this pattern of avoidance may qualify as "urgent" in any interesting sense of the term.

More radically, there are folk emotion categories that are borderline cases of umotions, and folk emotion categories that have no instances that qualify as umotions. The latter case may be that of regret and melancholy: they are perfectly legitimate folk emotions, but they most likely do not have any tokens that qualify as umotions. A more complex case is that of sadness and joy, which I discussed above in the context of Frijda's (1986) theory of emotions. If we understand urgent action tendencies as mechanisms for the urgent pursuit of specific relational goals, neither of them fits the bill. This is because they generally occur when a certain goal is either no longer achievable (sadness) or has already been achieved (joy). At the same time, if we think of control precedence in terms of global control on goal selection, some instances of sadness and joy may qualify as urgency management systems. This is because they, respectively, globally deactivate and globally activate the pursuit of an open class of relational goals. Whether or not this is a good way to include the cases of joy and sadness within the purview of the UMS theory is open for debate.

On the other hand, the notion of umotion comprises items which are not necessarily prototypical emotions, and possibly not even emotions, in the folk sense. An example may be that of "pain", which despite not being a prototypical emotion (and arguably not a folk emotion at all) appears to have instances that qualify as umotions. At first blush at least, "upain" is the urgent action tendency of recoiling and attempting to relieve a body part appraised as damaged, and it is PP appropriate in case real damage has been suffered.

Emotion theorists generally respond to the presence of vernacular emotions that fail to fit their favorite account (Type 1 counterexample), and vernacular non-emotions that fit it instead

(Type 2 counterexample) by trying to transform their defining notions so as to avoid all counterexamples. I could also start playing with the defining ingredients of the theoretical construct of "umotion" so as to try to encompass all and only those vernacular items we call "emotion" or "fear". For example, I could characterize "urgency" so broadly it that any form of action tendency, no matter how weak, counts as having control precedence if it is associated with an emotion. Under this account, even melancholy may count as an umotion.

I could further stipulate that pain is a bona fide emotion just because it fits my account, and that any instance of pain which is not an umotion is not "real" pain. In chapter 7, I documented how extensive the use of these ad hoc strategies is among cognitivists and Neo-Jamesians. Once we commit to the Explicating Emotion Project, and understand the ground rules of explication, these strategies quickly reveal their lack of theoretical payoff. There is nothing to gain for an explicative theory such as UMS theory in accommodating the cases of melancholy and regret, or excluding the case of pain, unless doing so reduces vagueness, increases fruitfulness or is required to achieve similarity in use with the explicandum. None of these desiderata is promoted by tweaking with the defining notions of umotion with the only purpose of capturing all and only those things we call emotion in ordinary language.

In order not to take the UMS theory seriously, a critic would have to argue either that it wears its lack of potential fruitfulness on its sleeves, or that it fails to achieve similarity in use with the ordinary notion of "emotion". I believe the UMS theory has the resources to counter both moves. Importantly, I do not propose the UMS theory as the only fruitful explication of folk emotion categories. Given the heterogeneity and vagueness of folk emotion categories, I believe there is no viable alternative in emotion theory to what I will call *explicative pluralism*. In my view, emotion theorists not interested in the project of capturing the family resemblance of folk emotion categories (Folk Emotion Project) have to explicate them in light of the theoretical objectives of their specific discipline. There will be many *explicata* "similar in use" to a given explicandum and fruitful relative to a given set of theoretical objectives, and there will be many theoretical objectives of different disciplines relative to which fruitful *explicata* for folk emotion categories can be generated. This approach will bring about a plurality of theoretical constructs which need not be considered in competition with one another, as long as they are offered in the context of the Explicating Emotion Project.

For example, affect program theorists would be mistaken if they thought that the UMS theory is an alternative to the notion of a "basic emotion" (see chapter 5). The notion of "basic emotion", which certainly differs from that of "umotion", has proven very fruitful in the sciences of mind. It was used to formulate many interesting inductive generalizations, explanations and predictions for example in biology and the neurosciences. Such generalizations concern the presence of homologies in facial expressions of basic emotions in other primates (e.g. Chevalier-Skolnikoff (1973) and across cultures (e.g. Ekman 1972), and homologies in the neural pathways of basic anger, basic fear and basic disgust (e.g. Lawrence and Calder 2004).

Umotion is just another potentially fruitful explicatum for the folk category of "emotion". Differently from the construct of "basic emotion", according to which nothing counts as a basic emotion unless it is automatically elicited, unbidden, short-lived, and with a distinctive physiology (see chapter 5), the construct of "umotion" is more suitable to shed light on the so-called "higher cognitive emotions", which generally lack such properties. Higher cognitive emotions such as guilt, shame, embarrassment, etc. can manifest strategic dimensions, they often last for a long time, and they generally lack a distinctive physiology.

Umotion may help us shed new light on the higher cognitive emotions by offering, among other things, a novel understanding of the importance of their communicative dimension and of their dynamic development through time. These aspects have largely been ignored by cognitivists and Neo-Jamesians, but they may hold the key to understanding the origin and current function of many higher cognitive emotions. Since these are the sorts of emotions involved in morality, art, mental disorder, etc., I am hopeful that the theoretical construct of "umotion" will help us shed light on such phenomena, which are those we are most eager to understand but know the least about. To see if the construct of "umotion" can do for the understanding of some forms of guilt, shame and embarrassment what the theoretical construct of "basic emotion" has done for the understanding of some forms of fear, anger, and disgust, what we need is to start using it in the context of theoretical projects involving such emotions. This is what I plan to do in the next few years of my academic career, hopefully in collaboration with scientists who will find the questions to which thinking of emotions as umotions leads us worth pursuing.

10.4. CONCLUSION

In this chapter, I have offered a new theory of emotions as urgency management systems, or umotions. At the heart of the theory I offered lies the idea that an "umotion" is a special type of superordinate system which instantiates and manages an urgent action tendency by coordinating the operation of a cluster of cognitive, perceptual and motoric subsystems. Crucially, such superordinate system has a proper function by virtue of which it acquires a special kind of intentionality I have called pragmatic. The fundamental idea here is that emotions combine descriptive and directive purposes, and rely on a system of representation which differs in kind from the one involved in the formation of beliefs and desires.

My theory differs very significantly from both cognitivist and Neo-Jamesianian theories, currently the two most popular accounts of emotions. They take the fundamental mark of the emotional to be, respectively, the evaluation embodied by an emotion and the way the emotion feels. I have instead constructed a theory of emotions around the idea that the fundamental mark of the emotional is urgency, understood as priority in the control of action. My theory is unlike cognitivism and Neo-Jamesianism also methodologically. I do not claim to have captured anything that deserves to be called an emotion in ordinary language. What I have claimed is instead that we must divorce the objectives of striving for ordinary language compatibility from the objective of striving for theoretical fruitfulness. Achieving the former is the project pursued by folk emotion theorists, and achieving the latter is the project pursued by theorists interested in explication. I am interested in scientific explication, and I have argued that the theoretical construct of umotion offers a good explication for folk emotion categories, one which will prove its fruitfulness if adopted by working scientists.

11. CONCLUSION

The debate between rival research programs in contemporary emotion theory is afflicted by two major problems which stand in the way of progress. The first is a lack of appreciation for the history of the subject, which has prevented many emotion theorists from learning from the insights and mistakes of their intellectual ancestors, often turned into anachronistic caricatures. The second is lack of methodological self-consciousness. Emotion theorists ask "What is an emotion?" without a clear understanding of what counts as getting the answer right. The first two parts of this dissertation have been devoted to trying to solve these problems, and at the same time prepare the way for a new theory of emotions built on historical understanding and sound methodology. I offered such theory in the third and last part of my dissertation.

In the historical part, I have distinguished between five main traditions that have battled for the soul of emotions in the past 2,500 years. I called them the feeling tradition, the cognitivist tradition, the behaviorist tradition, the evolutionary tradition and the social constructionist tradition. Studying a handful of central figures within each tradition has given me an opportunity to understand why emotions have been identified in the course of their long intellectual history with items as diverse as perceptions of bodily changes, judgments, behavioral predispositions, biologically based solutions to fundamental life tasks, and culturally specific social artifacts. This historical investigation has also revealed the existence of a significant area of agreement between rival traditions.

First, due largely to the efforts of cognitivists, the view that *emotions have intentionality* has gained wide currency in emotion theory. Secondly, there is now general agreement that at least some *emotions are shared by animals, infants and adult humans*. A third area of agreement concerns what we may call the *modularity properties of emotions*. It is fairly

uncontroversial at this stage that many emotions have a number of the characteristics of Fodorian modules. They are elicited at least some of the time by a mechanism which is fast and mandatory, and partially insulated from the one involved in the production and manipulation of linguistic representations. This insulation manifests itself in a variety of ways. For example, the operative principles of the eliciting mechanism are often unavailable to conscious report and impenetrable to beliefs, desires and other propositional attitudes.

The fourth area of agreement concerns the *basic components associated with emotions*, either necessarily or contingently. Consider an episode of intense anger directed by a scholar towards the colleague who has just informed her that she has been denied tenure. As a first approximation, we can distinguish in the complex event that is anger an *evaluative* component (e.g. appraising being denied tenure as a slight), a *physiological* component (e.g. increased heart rate and blood pressure), a *phenomenological* component (e.g. an unpleasant feeling), an *expressive* component (e.g. fixed stare, loud voice, erected body), a *behavioral* component (e.g. insulting, storming out of the room), a *mental* component (e.g. focusing attention, planning an appeal), and a *communicative* component (e.g. by getting angry the scholar conveys the message that she feels unjustly treated).

The question that has historically excited emotion theorists more than any other is: Which subset of the components I mentioned – evaluative, physiological, phenomenological, expressive, behavioral, mental, and communicative – is *essential* to emotion/anger? It is with respect to this question that different research programs part ways, and begin their long-running and often vicious theoretical disputes.

The methodological part of this dissertation has tried to expose such disputes as resulting in large part from confusion on the aims of theory construction. I argued that there are two importantly different ways of getting the answer right to a question of the form “What is an emotion?” or “What is anger?’. One is to capture the conditions of application of the folk term "emotion/anger" in ordinary language, and the other is to formulate a fruitful explication of it. These two objectives are equally legitimate but demand the application of different methodologies, neither of which appears to be implemented by contemporary emotion theorists. The main task of part two of my dissertation has been to clearly articulate the desiderata for what I have called the Folk Emotion Project and the Explicating Emotion Project.

On the basis of empirical evidence on folk emotion concepts, I have concluded that folk emotion categories are characterized by prototypical organization, blurred edges and extreme heterogeneity. Trying to capture the condition of membership of such categories with a definition, I argued, is an ill-conceived intellectual pursuit. There simply is no set of individually necessary and jointly sufficient conditions for something to count as either "emotion" or "anger" in the ordinary sense of such terms. Providing a cluster account of vernacular emotion categories seems to me the best an emotion theorist can ever do in the context of the Folk Emotion Project.

An emotion theorist engaged in explication, instead, must only achieve similarity in use between his favored explicatum and "emotion" or "anger", showing what useful theoretical purposes are served by it. Emotion theorists appear instead convinced that any good theory of emotions must encompass anything we call emotion in ordinary language, and that one and only one such theory can be found. This is because the desiderata of *theoretical fruitfulness* and *ordinary language compatibility* are not divorced in the mind of most emotion theorists. The implicit assumption is that what emotion terms "mean" in ordinary language coincides with what is a "good thing to mean" by them relative to the purposes of a scientific theory.

I argued that this is a bad assumption, because it is very unlikely that the items belonging to folk emotion categories share a scientifically interesting dimension of similarity. To do science, I concluded, an emotion theorist must go the way of explication. Since (a) explication aims to transform an explicandum category into an explicatum category similar to it but endowed with a higher degree of fruitfulness with respect to certain theoretical objectives, and (b) there are many ways to instantiate a similarity relation and be fruitful relative to sets of theoretical objectives, explication is intrinsically pluralistic. This insight has gone largely lost among contemporary emotion theorists, who argue as if there could only be one legitimate explication of emotions fruitful with respect to all conceivable theoretical objectives. But this is clearly a false assumption. This implies that many of the disputes in which contemporary emotion theorists are engaged are, once again, based on lack of methodological self-consciousness.

The constructive part of my dissertation has been devoted to the formulation of a new explication of emotion suitable for the theoretical purposes of scientific psychology. At the heart of the Urgency Management System (UMS) theory of emotions I proposed is the idea

that an “umotion” is a special type of superordinate system which instantiates and manages an urgent action tendency by coordinating the operation of a cluster of cognitive, perceptual and motoric subsystems. Crucially, such superordinate system has a proper function and a teleosemantic intentional content undifferentiated between presenting facts and directing activities appropriate to them. Just to give an example, fear does not independently tell us that danger is present and that some evasive action must be taken, but both things at the same time. I have parted ways with both cognitivists and Neo-Jamesianians, who take the mark of the emotional to be, respectively, the evaluation embodied by an emotion and the way the emotion feels, and argued instead that the fundamental mark of the emotional is urgency, understood as priority in the control of action.

The account I offered accommodates all the areas of agreement I singled out at the positive legacy of many centuries of investigation of the emotions, and it does not make the methodological mistakes so detrimental to many contemporary emotion theories.

The UMS theory explains in what sense umotions have intentionality, by offering an account of their pragmatic content and proper function. It explains why some umotions are shared by animals, infants and adult humans, as organisms of many different levels of cognitive sophistication can equally engage in urgent action tendencies governed by a mechanism with a history of selection. It accommodates the fact that some umotions have modularity properties, by allowing for appraisal, the main producer of umotions, to range from a modular to a central end. Finally, the UMS theory finds a place for all marks of emotionality, organizing evaluative, physiological, phenomenological, expressive, behavioral, mental, and communicative components in the context of three main functional components of preparation, action and communication. Umotions, I have argued, can be instantiated with a variety of different forms of preparation, action and communication.

Methodologically, I have been at pains to emphasize that the UMS theory is an *explication* of folk emotion categories, which does not have the ambition to capture the common ground shared by all things we call emotions in ordinary language. I am happy to concede that there are things we legitimately call “emotion” in English which are not “umotions”. At the same time, the notion of an urgency management system is similar enough in use to “emotion” to count as explicating it, it is defined fairly precisely, and, I have argued, it promises to be a fruitful notion for scientific psychology.

BIBLIOGRAPHY

Abu-Lughod, L. (1986). Veiled sentiments. Berkeley, University of California Press.

Allport, F. H. (1924). Social psychology. Boston, Houghton Mifflin.

Alston, W., P. (1967). Vagueness. The Encyclopedia of Philosophy. P. Edwards, Macmillan**:** 218-21.

Aristotle (1984). The Complete Works of Aristotle. Princeton, NJ, Princeton University Press.

Armon-Jones, C. (1985). "Prescription, explication & the social construction of emotions." Journal for the Theory of Social Behaviour **15 (1)**: 1-22.

Armon-Jones, C. (1986a). The thesis of constructionism. The Social Construction of Emotion, Harré,R (Ed) Oxford University Press**:** 32-56.

Armon-Jones, C. (1986b). The social functions of emotion. The Social Construction of Emotion, Harré, R. (Ed), Oxford, OUP**:** 57-82.

Armstrong, S., L. Gleitman, et al. (1983). "What some concepts might not be." Cognition **13**: 263-308.

Arnold, M. B. (1960). Emotion & Personality, Columbia Uni. Press.

Austin, J. L. (1962). How to do things with words. Cambridge, Harvard University Press.

Averill, J. R. (1980). Emotion & anxiety: sociocultural, biologicaland psychological determinants. Explaining Emotions, Rorty, A.O. (ed) University of California Press**:** 37-72.

Averill, J. R. (1986). The acquisition of emotions during adulthood. The Social Construction of Emotion, Harré, R (Ed) Oxford,OUP**:** 98-118.

Averill, J. R. (1975). "A semantic atlas of emotional concepts." JSAS: Catalog of Selected Documents in Psychology **5**(330).

Ax, A. F. (1953). "The physiological differentiation between fear and anger in humans." Psychosomatic Medicine **15 (5)**: 433-442.

Baltzly, D. (Winter 2004 Edition). "Stoicism." The Stanford Encyclopedia of Philosophy, from http:/plato.stanford.edu/archives/win2004/entries/stoicism/.

Bard, P. (1929). "The central representation of the sympathetic system: As indicated by certain physiological observations." Archives of Neurology and Psychiatry **22**: 230-46.

Barsalou, L. W., & Sewell, D. R. (1985). "Contrasting the representation of scripts and categories." Journal of Memory and Language **24**: 646-665.

Bateson, G., & Mead, M. (1942). Balinese character. New York, Academy of Sciences.

Bedford, E. (1957). " Emotions." Proceedings of the Aristotelian Society **57**: 281-304.

Bell, C. (1844). Anatomy and Philosophy of Expression as Connected with the Fine Arts. London, John Murray.

Bermúdez, J. L. (1998). The Paradox of Self-consciousness. Cambridge, MIT Press.

Berridge, K. C., & Winkielman, P. (2003). "What is an unconscious emotion: The case for unconscious 'liking'." Cognition and Emotion **17**: 181-211.

Birdwhistell, R. L. (1970). Kinesics and context. Philadelphia, University of Pennsylvania Press.

Block, N. J. (1995). "On a Confusion about a Function of Consciousness." Behavioral and Brain Sciences **18**: 227-47.

Block, N. J., O. J. Flanagan, et al. (1997). The nature of consciousness: philosophical debates. Cambridge, Mass., MIT Press.

Boakes, R. A. (1984). From Darwin to behaviourism: psychology and the minds of animals. Cambridge Cambridgeshire; New York, Cambridge University Press.

Boucher, J. D. and O. E. Carlson (1980). "Recognition of facial expression in three cultures." *Journal of Cross-cultural Psychology* **11**: 263-280.

Boyd, R. (1991). "Realism, anti-foundationalism and the enthusiasm for natural kinds." Philosophical Studies **61**: 127-148.

Brandom, R. (1994). Making it explicit: reasoning, representing, and discursive commitment. Cambridge, Mass., Harvard University Press.

Brandom, R. (2000). Articulating reasons: an introduction to inferentialism. Cambridge, Mass., Harvard University Press.

Briggs, J. L. (1970). Never in anger: Portrait of an Eskimo family. Cambridge, MA, Harvard University Press.

Broad, C. D. (1988). Emotions & sentiment. Broad's Crit. Essays in Moral Phil., Cheney (ed).

Brothers, L. (1997). Friday's blueprint: How society shapes the human mind. New York, Oxford University Press.

Buller, D. J. (2005). Adapting Minds: Evolutionary Psychology and the Persistent Quest for Human Nature, The MIT Press.

Buss, D. M. (2000). The Dangerous Passion: Why Jealousy is as Essential as Love and Sex. New York, Simon and Schuster.

Cacioppo, J. T., Berntson, G. G., Larsen, J. T., Poehlmann, K. M. & Ito, T. A. (2000). The Psychophysiology of Emotion. Handbook of Emotions. M. L. a. J. M. Haviland-Jones. New York, Guilford Publications.

Camras, L. A., Campos, J. J., Oster, H., Miyake, K., & Bradshaw, D. (1992). "Japanese and American infants responses to arm restraint." Developmental Psychology **28**: 578-583.

Cannon, W. (1929). Bodily Changes in Pain, Hunger, Fear and Rage. New York, Appleton.

Cantor, N., & Mischel, W. (1977). "Traits as prototypes: Effects on recognition memory." Journal of Personality and Social Psychology **35**: 38-48.

Cantor, N., Mischel, W., & Schwartz, J. C. (1982). "A prototype analysis of psychological situations." Cognitive Psychology **14**: 45-77.

Cantor, N., Smith, E. E., French, R. D., & Mezzich, J. (1980). "Psychiatric diagnosis as prototype categorization." *Journal of Abnormal Psychology* ***89***: 181-193.

Carnap, R. (1950). Logical foundations of probability. Chicago, University of Chicago Press.

Casati, R. and A. Varzi. (Fall 2002 Edition). "Events." The Stanford Encyclopedia of Philosophy, from http:/plato.stanford.edu/archives/fall2002/entries/events.

Charland, L. C. (2002). "The Natural Kind Status of Emotion." British Journal for the Philosophy of Science **53**: 511-537.

Chevalier-Skolnikoff, S. (1973). Facial expression of emotion in non-human primates. Darwin and Facial Expression: A Century of Research in Review. P. Ekman. New York and London, Academic Press**:** 11-89.

Chomsky, N. (1959). "Review of B.F. Skinner's "Verbal Behaviour"." Language **35**: 26-58.

Cooper, J. M. (1999). "Reason and Emotion."

Coulter, J. (1986). Affect and social context: emotion definition as a social task. The Social Construction of the Emotions, Harré.

Damasio, A. R. (1994). Descartes Error: Emotion, Reason and the Human Brain. New York, Grosset/Putnam.

Damasio, A. R. (1999). The Feeling of What Happens: Body and Emotion in the Making of Consciousness. New York, Harcourt Brace.

Damasio, A. R. (2003). Looking for Spinoza: joy, sorrow, and the feeling brain. Orlando, Fla., Harcourt.

Darwin, C. (1859/1968). The Origin of Species by Means of Natural Selection, or, the Preservation of Favoured Races in the Struggle for Life (facsimile edition). London, Penguin Books.

Darwin, C. (1871/1981). The Descent of Man and Selection in Relation to Sex. Princeton, NJ, Princeton University Press.

Darwin, C. (1872). The Expressions of Emotions in Man & Animals. New York, Philosophical Library.

Darwin, C. (1872/1955). The Expression of the Emotions in Man and Animals. New York, Philosophical Library.

Darwin, C. and P. Ekman (1997/1872). The expression of the emotions in man and animals. Oxford; New York, Oxford University Press.

Davidson, D. (1980). Essays on Actions and Events. Oxford, Clarendon Press.

Dawkins, R. and J. R. Krebs (1978). Animal signals: information or manipulation. Behavioral Ecology: An Evolutionary Approach. J. R. Krebs and N. B. Davies. Oxford, Blackwell.

De Sousa, R. (1987). The Rationality of Emotions. Cambridge, Mass., MIT Press.

Deigh, J. (1994). "Cognitivism in the theory of emotions." Ethics **104**: 824-854.

Delancey, C. (2001). Passionate Engines: What emotions reveal about mind and artificial intelligence. New York, Oxford, Oxford University Press.

Descartes, R. and G. Rodis-Lewis (1650/1955). Les passions de l'âme. Paris, J. Vrin.

Dewar, A. M. (1986). The social construction of gender in a physical education programme, University of British Columbia. **Master**.

Dretske, F. (1981). Knowledge and the Flow of Information. Oxford, Blackwells.

Dretske, F. (1986). Misrepresentation. Belief: Form, Content and Function. R. J. Bogdan. Oxford, Oxford University Press.

Dretske, F. (1988). Explaining Behaviour. Cambridge, MA, Bradford/MIT.

Ducci, L., Arcuri, L., Georgis, T. & Sineshaw, T. (1982). "Emotion recognition in Ethiopia." Journal of Cross-cultural Psychology **13**: 340-351.

Duchenne, G. B. (1990 (1867)). The Mechanism of Human Facial Expression. Cambridge, Cambridge University Press.

Eibl-Eibesfeldt, I. (1973). Expressive behaviour of the deaf & blind born. Social Communication & Movement. M. von Cranach and I. Vine. London & New York, Academic Press**:** 163-194.

Ekman, P. (1980). Biological & cultural contributions to body & facial movement in the expression of emotions. Explaining Emotions. A. O. Rorty. Berkeley, University of California Press**:** 73-102.

Ekman, P. (1992). "Facial Expressions of emotion: new findings, new questions." Psychological Science **3**: 34-38.

Ekman, P. (1994). "Strong Evidence for Universals in Facial Expressions: A Reply to Russell's Mistaken Critique." Psychological Bulletin **115(2)**: 268–287.

Ekman, P. (1999a). Facial Expressions. Handbook of Cognition and Emotion. T. Dalgleish and M. J. Power. Chichester, John Wiley and sons**:** 301-320.

Ekman, P. (1999b). Basic Emotions. Handbook of Cognition and Emotion. T. Dalgleish and M. Power. Chichester, John Wiley and Sons Co.**:** 45-60.

Ekman, P. (2003). Emotions Revealed. New York, Times Books.

Ekman, P. and W. V. Friesen (1969). "The repertoire of nonverbal behaviour." Semiotica **1 (1)**: 86-8.

Ekman, P. and W. V. Friesen (1971). "Constants across cultures in the face & emotion." Journal of Personality & Social Psychology **17 (2)**: 124-129.

Ekman, P. and W. V. Friesen (1986). "A new pan-cultural expression of emotion." *Motivation and Emotion* **10**: 159-168.

Ekman, P., W. V. Friesen, et al. (1987). "Universals and Cultural Differences in the Judgments of Facial Expresions of Emotion." Journal of Personality and Social Psychology **53(4)**(1987): 712-717.

Ekman, P., Friesen, W.V., & Hager, J.C. (2002). THE FACIAL ACTION CODING SYSTEM. Salt Lake City, Research Nexus eBoo.

Ekman, P. and K. O. Heider (1988). "The universality of contempt expression: a replication." *Motivation and Emotion* **12**: 303-308.

Ekman, P., R. W. Levenson, et al. (1983). "Autonomic nervous system activity distinguishes between emotions." Science **221**: 1208-1210.

Ekman, P., Sorenson, E. R. & Friesen. W. V. (1969b). "Pan-cultural elements in facial displays of emotions." Science **164**(3875): 86-88.

Fehr, B. and J. A. Russell (1984). "Concept of emotion viewed from a prototype perspective." Journal of Experimental Psychology: General **113**: 464-86.

Fernández-Dols, J. M. and J. M. Carrol (1997a). Is the meaning percieved in facial expression independent of its context? The Psychology of Facial Expression. J. A. Russell and J. M. Fernández-Dols. Cambridge, Cambridge University Press**:** 295-320.

Fernández-Dols, J. M. and M.-A. Ruiz-Belda (1997b). Spontaneous facial behavior during intense emotional episodes: Artistic truth and optical truth. The Psychology of Facial Expression. J. A. Russell and J. M. Fernández-Dols. Cambridge, Cambridge University Press**:** 255-294.

Fodor, J. A. (1983). The Modularity of Mind: An Essay in Faculty Psychology. Cambridge, Mass, Bradford Books/MIT Press.

Foley, J. P., Jr. (1935). "Judgment of facial expression of emotion in the chimpanzee." Journal of Social Psychology **6**: 31-54.

Freud, S. (1915). The Unconscious. On Metapsychology: the Theory of Psychoanalysis. Beyond the Pleasure Principle, The Ego and the Id, and other works. Middlesex, Pelican Books**:** 159-222.

Freud, S. (1923). The Ego and the Id. On Metapsychology: the Theory of Psychoanalysis. Beyond the Pleasure Principle, The Ego and the Id, and other works. Middlesex, Pelican Books**:** 339-407.

Freud, S. (1940). An Outline of Psychoanalysis. Historical and Expository Works on Psychoanalysis: History of the Psychoanalytic Movement, An Autobiographical Study, Outline of Psychoanalysis, and other works. London, Penguin Books**:** 369-443.

Fridlund, A. (1997). The new ethology of human facial expressions. The Psychology of Facial Expressions. J. A. Russell and J. M. Fernández-Dols. Cambridge, Cambridge University Press**:** 103-129.

Fridlund, A. J. (1989). Evolution and facial action in reflex, social motive, and paralanguage, University of California.

Fridlund, A. J. and B. Duchaine (1996). "Facial expressions of emotions" and the delusion of the hermetic self. The emotions: social, cultural and biological dimensions. R. Harré and W. G. Parrott. London; Thousand Oaks, Calif., Sage Publications**:** viii, 323.

Fridlund, A. J., J. A. Schaut, et al. (1990). "Audience effects on solitary faces during imagery: displaying to the people in your head." Journal of Nonverbal Behaviour **14(2)**: 113-137.

Frijda, N. H. (1986). The Emotions. Cambridge, Cambridge University Press.

Geertz, Z. H. (1959). "The vocabulary of emotion: A study of Javanese socialization processes." Psychiatry **22**: 225-237.

Gerber, E. (1975). The cultural patterning of emotions in Samoa. San Diego, University of California, San Diego. **PhD**.

Gerber, E. R. (1985). Rage and obligation: Samoan emotion in conflict. Person, self and experience: Exploring Pacific ethnopsychologies. G. M. W. J. Kirkpatrick. Berkeley, University of California Press**:** 121-167.

Gibbard, A. (1990). Wise Choices, Apt Feelings: A Theory of Normative Judgment. Cambridge, Mass., Harvard University Press.

Gibson, J. J. (1979). The Ecological Approach to Visual Perception. Boston, Houghton Mifflin.

Goldie, P. (2004). Emotion, Feeling, and Knowledge of the World. Thinking About Feeling: Contemporary Philosophers on Emotions. R. C. Solomon.

Goodenough, F. L. (1931). "The expression of the emotions in infancy." Child Development, **2**: 96-101.

Gordon, R. M. (1987). The structure of emotions: investigations in cognitive philosophy. Cambridge [Cambridgeshire]; New York, Cambridge University Press.

Gould, S. J. and R. Lewontin (1978). "The Spandrels of San Marco and the Panglossian Paradigm: a critique of the adaptationist programme." Proceedings of the Royal Society of London B **205**: 581-598.

Greenspan, P. (1988). Emotions and Reasons:An Inquiry into Emotional Justification. New York, Routledge.

Griffiths, P. E. (1997). What Emotions Really Are: The Problem of Psychological Categories. Chicago, University of Chicago Press.

Griffiths, P. E. (2003). Philosophy and the emotions. Royal Institute of Philosophy supplement; 52. A. Hatzimoysis. New York, Cambridge University Press**:** 39-67.

Griffiths, P. E. (2004a). Towards a Machiavellian theory of Emotional Appraisal. Emotion, Evolution and Rationality. P. Cruise and D. Evans. Oxford, Oxford University Press**:** 89-105.

Griffiths, P. E. (2004b). Is emotion a natural kind? Thinking About Feeling: Contemporary Philosophers on Emotions. R. C. Solomon. Oxford, Oxford University Press**:** 233-249.

Grünbaum, A. (1984). The Foundations of Psychoanalysis. Berkeley, University of California Press.

Guzeldere, G. (1997). The Many Faces of Consciousness: a Field Guide. The nature of consciousness: philosophical debates. N. J. Block, O. J. Flanagan and G. Gèuzeldere. Cambridge, Mass., MIT Press.

Hacking, I. (1999). The social construction of what? Cambridge, Mass, Harvard University Press.

Haidt, J. (2001). "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." Psychological Review **118**(4): 814-834.

Hampton, J. A. (1979). "Polymorphous concepts in semantic memory." Journal of Verbal Learning and Verbal Behavior **18**: 441-461.

Harré, R. (1986). An outline of the social constructionist viewpoint. The Social Construction of Emotion, Harré, R (Ed) Oxford,OUP**:** 2-14.

Hartley, J., & Homa, D. (1981). "Abstraction of stylistic concepts." *Journal of Experimental Psychology: Human Learning and Memory* **7**: 33-46.

Helmholtz, H. v. (1860/1979). Treatise on Physiological Optics, vol. 3. Basic Writings in the History of Psychology. R. I. Watson. New York, Oxford University Press.

Hess, W. R. (1954). Dar Zwischenhirn. Basel, Schwabe.

Hiatt, L. R. (1978). Classification of the emotions. Australian aboriginal concepts. L. R. Hiatt. Princeton, NJ, Humanities Press**:** 182-187.

Hinde, R. A. (1985a). Expression and Negotiation. The Development of Expressive Behavior. G. Zivin. New York, Academic Press**:** 103-116.

Hinde, R. A. (1985b). "Was 'The Expression of Emotions' a misleading phrase?" Animal Behaviour(33): 985-992.

Howell, S. (1981). Rules not words. Indigenous psychologies: The anthropology of the self. P. H. A. Lock. San Diego, CA, Academic Press**:** 133-143.

Hume, D. (1739/1992). A treatise of human nature. Buffalo, Prometheus Books.

Hunsperger, R. W. (1956). Helvet. phvsiol. et pharmacol. acta 14 **70**.

Hurley, S. (2003). "Animal actions in the space of reasons." Mind and Language **18**(3): 231-256.

Izard, C. (1994). "Innate and Universal Facial Expressions: Evidence From Developmental and Cross-Cultural Research." Psychological Bulletin **115(2)**: 288-299.

Izard, C. E. (1969). The emotions & emotions constructs in personality and culture. Handbook of Modern Personality Theory. R. B. Cattell and R. M. Dreger. Washington, DC & London, Hemisphere Publishing Corporation.

Izard, C. E. (1971). The Face of Emotion. New York, Appleton, Century, Crofts.

Izard, C. E. (1977). Human Emotions. New York, Plenum.

Izard, C. E. (1980). Cross cultural perspectives on emotion and emotion communication. Handbook of cross-cultural psychology. Vol.3.Triandis, H & Lonner, W (eds). Allyn & Bacon, Boston.**:** 185-221.

Izard, C. E. (1992). "Basic emotions, relations amongst emotions and emotion-cognition relations." Psychological Review **99 (3)**: 561-565.

Izard, C. E. (1993). "Four systems for emotion activation: cognitive and noncognitive processes." Psychological Review **100**: 68-90.

James, W. (1884). "What is an emotion?" Mind **IX**: 188-205.

James, W. (1890). The Principles of Psychology. NY, Holt.

James, W. (1894). "The physical basis of emotion." Psychological Review **1**: 516-529.

Johnson, A., Johnson, O., & Baksh, M. (1986). "The colours of emotions in Machiguenga." American Anthropologist **88**: 674-681.

Keeler, W. (1983). "Shame and stage fright in Java." Ethos **11**: 152-165.

Keltner, D. (1995). "Signs of appeasement: evidence for the distinct displays of embarrassment, amusement, and shame." Journal of Personality and Social Psychology **68**: 441-454.

Kenny, A. (1963). Action, Emotion & Will. London, Routledge & Kegan Paul.

Kihlstrom, J. F. (1987). "The cognitive unconscious." Science **237**: 1445-52.

Kihlstrom, J. F., Mulvaney, S., Tobias, B.A., & Tobis, I.P. (1998). The emotional unconscious. Counterpoints: Cognition and emotion. J. F. K. E. Eich, G.H. Bower, J.P. Forgas, & P.M. Niedenthal. New York, Oxford University Press.

Kitcher, P. (2001). Battling the undead: How (and how not) to resist genetic determinism. Thinking about Evolution: Historical, Philosophical and Political Perspectives (Festchrifft for Richard Lewontin). R. Singh, K. Krimbas, D. Paul and J. Beatty. Cambridge, Cambridge University Press**:** 396-414.

Kivy, P. and P. Kivy (1989). Sound sentiment: an essay on the musical emotions, including the complete text of The Corded shell. Philadelphia, Temple University Press.

Klineberg, O. (1940). Social psychology. New York, Holt.

Kluver, H., and Bucy, P. C (1937). ""Psychic blindness" and other symptoms

following bilateral temporal lobectomy in rhesus monkeys." American Journal of Physiology **119**: 352-53.

Konner, M. (1982). The Tangled Wing: Biological Constraints on the Human Spirit. London, William Heinemann Ltd.

Kukla, A. (2000). Social Constructivism and the Philosophy of Science. London, Routledge.

Kundera, M. (1980). The book of laughter and forgetting. New York, Knopf.

Landis, C. (1924). "Studies of emotional reactions: II. General behavior and facial expression." Journal of Comparative Psychology **4**: 447-509.

Lange, C. (1885/1922). The emotions. Baltimore, Williams & Wilkins.

Latour, B. and S. Woolgar (1979). Laboratory Life: The Social Construction of Scientific Facts. Beverly Hills, Sage.

Laver, S. a. H., M. (1997). Emotion and the Arts. Oxford, Oxford University Press.

Lawrence, A. D. and A. J. Calder (2004). Homologizing human emotions. Emotion, Evolution and Rationality. D. E. a. P. Cruse. Oxford, Oxford University Press.

Lazarus, R. (2001). Relational meaning and discrete emotions. Appraisal Processes in Emotion. K. Scherer, R, Schorr, A., Johnstone, T. Oxford, Oxford University Press.

Lazarus, R. S. (1991). Emotion and Adaptation. New York, Oxford University Press.

Lazarus, R. S., & Baker, R. W (1956a). "Personality and psychological stress: A theoretical and methodological framework." Psychological Newsletter(8): 21-32.

Lazarus, R. S., & Baker, R. W. (1956b). "Psychology. Progress in Neurology and Psychiatry." **11**: 253-271.

Lazarus, R. S., Deese, J., & Osler, S. F. (1952). "The effects of psychological stress upon performance." Psychological Bulletin **49**: 293-317.

Leary, M. R., Landel, J. L., & Patton, K. M. (1996). "The motivated expression of embarrassment following a self-presentational predicament." Journal of Personality and Social Psychology **64**: 619-636.

Lebra, T. S. (1983). "Shame and guilt: A psychocultural view of the Japanese self." Ethos **11**: 192-209.

LeDoux, J. (1996). The Emotional Brain: The Mysterious Underpinnings of Emotional Life. New York, Simon and Schuster.

Leff, J. (1973). "Culture and the differentiation of emotional states." British Journal of Psychiatry **123**: 299-306.

Levenson, R. W., P. Ekman, et al. (1990). "Voluntary facial expression generates emotion-specific nervous system activity." Psychophysiology **27**: 363-384.

Levenson, R. W., P. Ekman, et al. (1992). "Emotion and autonomic nervous system activity in the Minangkabau of West Sumatra. Journal of Personality and Social Psychology." Journal of Personality and Social Psychology **62** (972-88).

Leventhal, H., & Scherer, K. R. (1987). "The relationship of emotion to cognition: A functional approach to a semantic controversy." Cognition and Emotion **1**: 3-28.

Levine, N. E. (1988). The dynamics of polyandry: Kinship, domesticity, and population on the Tibetan border. Chicago, University of Chicago Press.

Levy, R. I. (1973). Tahitians. Chicago, University of Chicago Press.

Levy, R. I. (1983). "Introduction: Self and emotion." Ethos **11**: 128-134.

Levy, R. I. (1984). The emotions in comparative perspective. Approaches to emotion. K. R. S. P. Ekman. Hillsdale, NJ, Erlbaum**:** 397-412.

Lorenz, K. (1965). Preface to 'The Expression of the Emotions in Man and Animals' by Charles Darwin. Chicago, University of Chicago.

Luke, C. (1990). Constructing the Child Viewer: A History of the American Discourse on Television and Children, 1950-1980. New York, Praeger.

Lutz, C. (1980). Emotion words and emotional development on Ifaluk Atoll, Harvard University. **PhD**.

Lutz, C. (1983). "Parental goals, ethnopsychology, and the development of emotional meaning." Ethos **11**: 246-262.

Lutz, C. (1985). Ethnopsychology compared to what? Explaining behaviour and consciousness among the Ifaluk. Person, self and experience: Exploring Pacific ethnopsychologies. G. M. W. J. Kirkpatrick. Berkeley, University of California Press**:** 35-79.

Lutz, C. (1988). Unnatural emotions: everyday sentiments on a Micronesian atoll & their challenge to western theory. Chicago, University of Chicago Press.

Machamer, P. and L. Osbeck (2003). Perception, Conception and the Limits of the Direct Theory. The Philosophy of Majorie Grene (Library of Living Philosophers). R. E. Auxier and L. E. Hahn. Peru, IL, Open Court.

Mackie, J. L. (1965). "Causes and Conditions." American Philosophical Quarterly **2**: 245-64.

MacLean, P. D. (1949). "Psychosomatic disease and the "visceral brain": recent developments bearing on the Papez theory of emotion." Psychosomatic Medicine **11**: 338-53.

MacLean, P. D. (1952). "Some psychiatric implications of physiological studies on frontotemporal portions of the limbic system (visceral brain)." Electroencephlography & Clinical Neurophysiology **4**: 407-418.

MacLean, P. D. (1960). Psychosomatics. Handbook of Physiology, Vol. III, American Physiologicsal Society, Washington**:** 1723-44.

MacLean, P. D. (1969). The hypothalamus & emotional behaviour. The Hypothalamus, Haymaker, W. Schmidt, F.O et al (Eds) NY, Rockerfeller.

Malamud, N. (1966). "The epileptogenic focus in temporal lobe epilepsy from a pathological standpoint." Arch Neurol. **14**: 190-195.

Malmo, R. B., Shagass, C., & Davis, F. H. (1950). "Symptom specificity and bodily reactions during psychiatric interview." Psychosomatic Medicine **12**: 362-376.

Marcel, A. J. (1983). "Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes." Cognitive Psychology(15): 238-300.

Markus, H. R. and S. Kitayama (1991). "Culture and the self: Implications for cognition, emotion, and motivation." Psychological Review **98**: 224-253.

Marr, D. (1982). Vision. New York, W.H. Freeman.

Marsella, A. J. (1981). Depressive experience and disorder across cultures. Handbook of cross-cultural psychology. H. T. J. Draguns. Boston, Allyn & Bacon**:** 237-289.

Matravers, D. (2001). Art and Emotion. Oxford, Oxford University Press.

Matsumoto, D. R. (1992). "More evidence for the universality of a contempt expression." Motivation and Emotion **16**: 363-368.

McAndrew, F. T. (1986). "A cross-cultural study of recognition thresholds for facial expression of emotion." Journal of Cross-cultural Psychology **17**: 211-224.

McCormick, C. (1995). Constructing Danger: The Mis/representation of Crime in the News. Halifax, N. S., Fernwood.

Mesquita, B., & Frijda, N. H. (1992). "Cultural variations in emotions: A review." Psychological Bulletin **112**: 179-204.

Miceli, M. (1992). "How to make someone feel guilty. Strategies of guilt inducement and their goals." Journal for the Theory of Social Behavior **22**: 81-104.

Millikan, R. G. (1984). Language, thought, and other biological categories: new foundations for realism. Cambridge, Mass., MIT Press.

Millikan, R. G. (1993). White Queen psychology and other essays for Alice

Ruth Garrett Millikan. Cambridge, Mass., MIT Press.

Millikan, R. G. (1996). Pushmi-pullyu representations. Philosophical Perspectives. J. Tomberlin. Atascadero, CA, Ridgeview. **Vol. 9:** 185-200.

Millikan, R. G. (2000). On clear and confused ideas: an essay about substance concepts. Cambridge England; New York, Cambridge University Press.

Millikan, R. G. (2004). Varieties of meaning: the 2002 Jean Nicod lectures. Cambridge, Mass., MIT Press.

Modrak, D. (1983). Forms and Compounds. How Things Are: Studies in Predication and the History of Philosophy. J. Bogen and J. E. McGuire. Dordrecht, Reidel**:** 85-99.

Morice, R. (1978). "Psychiatric diagnosis in a transcultural setting: The importance of lexical categories." *British Journal of Psychiatry* ***132***: 87-95.

Moussa, H. (1992). The Social Construction of Women Refugees: A Journey of Discontinuities and Continuities, University of Toronto. **Ed. D.**

Mumford, S. (1998). Dispositions. Oxford, Oxford University Press.

Murphy, G. L. (2003). The Big Book of Concepts. Cambridge, MA, MIT Press.

Newman, P. L. (1964). ""Wild man" behaviour in a New Guinea Highlands community." American Anthropologist **66**: 1-19.

Niit, T. V., J. (1991). "Recognition of facial expressions: an experimental investigation of Ekman's model." Acta et Commentationes Universitatis Tarvensis **429**: 85-107.

Nussbaum, M. C. (2001). Upheavals of thought: the intelligence of emotions. Cambridge, New York, Cambridge University Press.

O'Malley, M. N., & Greenberg, J. (1983). "Sex differences in restoring justice: The down payment effect." Journal of Research in Personality **17**: 174-185.

Obeyesekere, G. (1981). *Medusa's hair: An essay on personal symbols and religious experience*. Chicago, University of Chicago Press.

Öhman, A. (1999). Distinguishing Unconscious from Conscious Emotional Processes: Methodological Considerations and Theoretical Implications. Handbook of Emotion and Cognition. T. Dalgleish and M. J. Power. Chichester, John Wiley and sons**:** 321-352.

Öhman, A. (2002). "Automaticity and the amygdala: Nonconscious responses to emotional faces." Current Directions in Psychological Science **11**(2): 62-66.

Oster, H., Hegley. D. & Nagel, L. (1992). "Adult judgment and fine-grained analysis of infant facial expression: testing the validity of a priori coding formulas." Developmental Psychology **28**: 1115-1131.

Panksepp, J. (1998). Affective Neuroscience: The Foundations Of Human And Animal Emotions. Oxford, Oxford University Press.

Panksepp, J. (2000). Emotions as natural kinds within the mammalian brain. Handbook of Emotions. M. Lewis and J. M. Haviland-Jones. New York and London, Guildford Press**:** 137-156.

Papez, J., W. (1937). "A proposed mechanism of emotion." Archives of Neurology and Psychiatry **79**: 217-24.

Parkinson, B. (1995). Ideas and Realities of Emotion. London and New York, Routledge.

Parkinson, B. (2004). Unpicking reasonable emotions. Emotion, Evolution and Rationality. P. Cruise and D. Evans. Oxford, Oxford University Press**:** 107-129.

Parkinson, B., A. Fischer, et al. (2005). Emotion in social relations: cultural, group, and interpersonal processes. New York, Psychology Press.

Penfield, W., Jaspers, H. & McNaughton, F. (1954). Epilepsy and the Functional Anatomy of the Human Brain. Boston, Little, Brown.

Pickering, A. (1984). Constructing Quarks: A Sociological History of Particle Physics. Edinburgh, Edinburgh University Press.

Pitcher, G. (1965). "Emotion." Mind **74**(285): 326-346.

Plutchik, R. (1962). The Emotions - Facts, Theories & A New Model. New York, Random House.

Plutchik, R. (1970). Emotions, evolution & adaptive processes. Feelings & Emotions. M. B. Arnold. New York, Academic Press**:** 3-24.

Plutchik, R. K., H., Ed. (1980). Emotion: Theory, Research and Experience Theories of Emotion. NY, Academic Press.

Poole, F. J. P. (1985). Coming into social being: Cultural images of infants in Bimin-Kuskusmin folk psychology. Person, self, and experience: Exploring Pacific ethnopsychologies. G. M. W. J. Kirkpatrick. Berkeley, University of California Press**:** 183-242.

Posner, M. I. (1978). Chronometric explorations of mind: the third Paul M. Fitts lectures, delivered at the University of Michigan, September 1976. Hillsdale, N.J.

New York, L. Erlbaum Associates.

Prinz, J. (2004a). Embodied emotions. Thinking About Feeling: Contemporary Philosophers on Emotions. R. C. Solomon. Oxford, Oxford University Press**:** 89-105.

Prinz, J. (2004b). Gut Reactions: A Perceptual Theory of Emotion. Oxford, Oxford University Press.

Prinz, J. (2004c). Which emotions are basic? Emotion, Evolution and Rationality. P. Cruise and D. Evans. Oxford, Oxford University Press**:** 69-87.

Prior, E. (1985). Dispositions. Aberdeen, Aberdeen University Press.

Redican, W. (1982). An Evolutionary Perspective on Human Facial Displays. Emotion in the Human Face. P. Ekman.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). "Semantic distance and the verification of semantic relations." Journal of Verbal Learning and Verbal Behavior **12**: 1-20.

Rorty, A. (1980). Essays on Aristotle's ethics. Berkeley, University of California Press.

Rosaldo, M. Z. (1980). Knowledge and passion: Ilongot notions of self and social life. Cambridge, England, Cambridge University Press.

Rosaldo, M. Z. (1983). "The shame of headhunters and the autonomy of self." Ethos **11**: 135-151.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. Cognitive development and the acquisition of language. T. E. Moore. San Diego, CA, Academic Press**:** 111-144.

Rosch, E. (1978). Principles of Categorization. Concepts: Core Readings. E. Margolis and S. Laurence. Cambridge, Mass., MIT Press**:** 189-206.

Rosch, E. and C. B. Mervis (1975). "Family resemblances: studies in the internal structure of categories." Cognitive Psychology **7**: 573-605.

Rosch, E., Simpson, C., & Miller, R. S. (1976). "Structural bases of typicality effects." Journal of Experimental Psychology: Human Perception and Performance, **2**: 491-502.

Rosenberg, E., & Ekman, P. (1993). *Conceptual and methodological issues in the judgment of facial expression of emotion*.

Russell, J. (1994). "Is There Universal Recognition of Emotion From Facial Expression? A Review of the Cross-Cultural Studies." Psychological Bulletin **115(1)**: 102-141.

Russell, J. A. (1991). "Culture and the categorization of emotions." Psychological Bulletin **110**: 426-450.

Russell, J. A. (1997). Reading emotion from and into faces: resurrecting a dimensional-contextual perspective. The Psychology of Facial Expression. J. A. Russell and J. M. Fernández-Dols. Cambridge, Cambridge University Press**:** 295-320.

Russell, J. A., & Fehr, B. (1994). "Fuzzy concepts in a fuzzy hierarchy: Varieties of anger." Journal of Personality and Social Psychology **67**: 186-205.

Russell, J. A., Bachorowski, J.-A., & Fernández-Dols, J. M. (2003). "Facial and vocal expressions of emotions." Annual Review of Psychology **54**: 349-359.

Ryle, G. (1949). The Concept of Mind. London, Hutchinson.

Sabini, J., & Silver, M. (1982). Moralities of everyday life. Oxford, England, Oxford University Press.

Sartre, J. P. and B. Frechtman (1948). The emotions, outline of a theory. New York, Philosophical Library.

Scarantino, A. (2004). "Affordances Explained." Philosophy of Science **71 (5)**(Supplement: Proceedings of the 2002 Biennial Meeting of the Philosophy of Science Association. Part I: ContributedPapers): 949-961.

Scarantino, A. (2005). Did Dretske Learn the Right Lesson from Shannon's Theory of Information? ms.

Schachter, S. and J. E. Singer (1962). "Cognitive, social and physiological determinants of emotional state." Psychological Review **69**: 379-399.

Schaffner, K. F. (1998). Chemical Systems and Chemical Evolution: the Philosophy of Molecular Biology. Philosophy of Biology. M. Ruse. Amherst, NY, Prometheus Books**:** 198-208.

Scherer, K. (2001). Appraisal considered as a process of multilevel sequential checking. Appraisal Processes in Emotion. K. Scherer, R, Schorr, A., Johnstone, T. Oxford, Oxford University Press.

Scherer, K. R., A. Schorr, et al. (2001). Appraisal processes in emotion: theory, methods, research. Oxford; New York, Oxford University Press.

Scherer, K. R., Wallbott, H. G., Matsumoto, D., & Kudoh, T. (1988). Emotional experience in cultural context: A comparison between Europe, Japan, and the United States. Facets of emotions. K. R. Scherer. Hillsdale, NJ, Erlbaum**:** 5-30.

Schieffelin, E. D. (1983). "Anger and shame in the tropical forest: An affect as a cultural system in Papua New Guinea." Ethos **11**: 181-191.

Schiffrin, R. M., & Schneider, W. (1977). "Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory." Psychological Review **84**: 127-190.

Searle, J. R. (1969). Speech acts: an essay in the philosophy of language. London, Cambridge U.P.

Searle, J. R. (1983). Intentionality, an essay in the philosophy of mind. Cambridge [Cambridgeshire]; New York, Cambridge University Press.

Searle, J. R., F. Kiefer, et al. (1980). Speech act theory and pragmatics. Dordrecht, Holland; Boston

Hingham, MA, D. Reidel;

sold and distributed in the U.S.A. and Canada by Kluwer Boston.

Seligman, M. (1971). "Phobias & preparedness." Behaviour Therapy **2(3)**: 307-320.

Sellars, W. (1966). Fatalism and Determinism. Freedom and Determinism. K. Lehrer. New York, Random House.

Semin, G. R. and A. S. R. Manstead (1982). "The social implications of embarrassment displays and restitution behavior." European Journal of Social Psychology **12**: 367-377.

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). "Emotion and emotion knowledge: Further explorations of a prototype approach." Journal of Personality and Social Psychology(52): 1061-1086.

Shostak, M. (1983). *Nisa: The life and words of a!Kung woman*. New York, Vintage.

Shweder, R. A. (1991). Thinking through cultures. Cambridge, MA, Harvard University Press.

Simon, H. A. (1967). "Motivational and emotional controls of cognition." Psychological Review(74): 29-39.

Skinner, B. F. (1953). Science and human behavior. New York, Macmillan.

Smith, E. E. (1989). Concepts and induction. *Foundations of cognitive science*. M. I. Posner. Cambridge, MA, MIT Press**:** 501-526.

Solomon, R. (1976). The Passions. New York, Doubleday.

Solomon, R. C. (2003). Not passion's slave: emotions and choice. Oxford; New York, Oxford University Press.

Sorabji, R. (2000). Emotion and Peace of Mind. Oxford, Oxford University Press.

Sorensen, R. (Fall 2003 Edition). "Vagueness." The Stanford Encyclopedia of Philosophy, from URL = <http://plato.stanford.edu/archives/fall2003/entries/vagueness/>.

Sorenson, E. R. (1976). The edge of the forest: Land, childhood and change in a New Guinea protoagricultural society. Washington, DC, Smithsonian Institution Press.

Speisman, J. C., Lazarus, R. S., Mordkoff, A., & Davison, L. (1964). "Experimental reduction of stress based on ego-defense theory." Journal of Abnormal and Social Psychology **68**: 367-380.

Stein, N. L., T. Trabasso, et al. (1993). The representation and organization of emotional exerience: unfolding the emotion episode. Handbook of Emotions. M. Lewis and J. M. Haviland. New York, Guildford Press**:** 279-300.

Stokes, A. W. (1962). "The comparative ethology of great, blue, marsh and coal tits at a winter feeding station." Behaviour **19**: 208-218.

Tangney, J. P. and K. W. Fischer, Eds. (1995). Self-Conscious Emotions: Shame, Guilt, Embarrassment, and Pride. New York, Guildford.

Tangney, J. P., Wagner, P. E., Hill-Barlow, D., Marshall, D. E., & Gramzow, R. (1996). "Relation of shame and guilt to constructive versus destructive responses in anger across the lifespan." Journal of Personality and Social Psychology **70**: 797-809.

Thompson, J. (1941). "Development of facial expressions of emotion in blind and seeing children." Archives of Psychology **37 (264)**: 1-47.

Tomkins, S. S. (1962). Affect, Imagery and Consciousness. New York, Springer.

Tomkins, S. S. and E. V. Demos (1995). Exploring affect: the selected writings of Silvan S. Tomkins. Cambridge England; New York

Paris, Cambridge University Press;

Editions de la Maison des sciences de l'homme.

Tooby, J. and L. Cosmides (1990). "The past explains the present: emotional adaptations and the structure of ancestral environments." Ethology and Sociobiology **11**: 375-424.

Tooby, J. and L. Cosmides (2000). Evolutionary psychology and the emotions. Handbook of Emotions. M. Lewis and J. M. Haviland-Jones. New York and London, Guildford Press**:** 91-116.

Tousignant, M. (1984). "Pena in the Ecuadorian Sierra: A psychoanthropological analysis of sadness." Culture, Medicine and Psychiatry **8**: 381-398.

Tye, M. (1996). Ten problems of consciousness: a representational theory of the phenomenal mind. Cambridge, Mass., MIT Press.

van Hooff, J. A. R. A. M. (1967). Facial displays of Catarrhine monkeys and apes. Primate ethology. D. Morris. London, Weidenfield & Nicholson**:** 7-68.

van Wyhe, J. "The Writings of Charles Darwin on the Web." Retrieved August, 2005, from http://pages.britishlibrary.net/charles.darwin/.

Vangelisti, A. L., Daly, J.A., & Rudnick, J.R. (1991). "Making people feel guilty in conversations: Techniques and correlates." Human Communication Research **18**: 3-39.

Wallace, A. F. C., & Carson, M. T. (1973). "Sharing and diversity in emotion terminology." Ethos **1**: 1-29.

Wallbott, H. G., & Scherer, K. R. (1988). How universal and specific is emotional experience? Evidence from 27 countries. K. R. Scherer. Hillsdale, N.J., Erlbaum**:** 31-56.

Watson, J. B. (1913). "Psychology as the Behaviorist Views it." Psychological Review **20**: 158-177.

Watson, J. B. (1919). Psychology from the standpoint of a behaviorist. Philadelphia, Lippincott.

Watson, J. B. (1925). Behaviorism. New York, Harpers.

Weiskrantz, L. (1956). "Behavioral changes associated with ablation of the amygdaloid complex in monkeys." Journal of Comparative Physiological Psychology **49**: 381-91.

Welker, R. L. (1982). "Abstraction of themes from melodic variation." Journal of Experimental Psychology: Human Perception and Performance **8**: 435-447.

Williamson, T. (1994). Vagueness. London, Routledge.

Witt, C. (1987). "Hylomorphism in Aristotle." Journal of Philosophy **84**: 673-679.

Wittgenstein, L. (1953). Philosophical Investigations (2nd ed). Oxford, Blackwell.

Woodfield, A. (1976). Teleology. Cambridge, Cambridge University Press.

Woodmansee, M. and P. Jaszi, Eds. (19994). The Construction of Authorship: Textual Appropriation in Law and Literature. Durhma, N.C., Duke University Press.

Wright, C. (2001). "On Being in a Quandary." Mind **110**: 45-98.

Wundt, W. (1896). Grundriss der Psychologie [Outlines of psychology]. Leipzig, Germany, Entgelmann.

Zajonc, R. (2000). Feeling and thinking: closing the debate on the independence of affect. Feeling and thinking: the role of affect in social cognition. Cambridge, U.K.; New York

Paris, Cambridge University Press**:** xvi, 421.

Zajonc, R. B. (1980). "Feeling & thinking: preferences need no inference." American Psychologist **35**: 151-175.