Reasons: A Naturalistic Explanation

Jurjako, Marko

UNIVERSITY OF RIJEKA

FACULTY OF HUMANITIES AND SOCIAL SCIENCES

Marko Jurjako

Reasons: A Naturalistic Explanation

DOCTORAL THESIS

Rijeka, 2016.

UNIVERSITY OF RIJEKA

FACULTY OF HUMANITIES AND SOCIAL SCIENCES

Marko Jurjako

Reasons: A Naturalistic Explanation

DOCTORAL THESIS

Advisor: prof. dr. sc. Nenad Smokrović

Co-advisor: doc. dr. sc. Luca Malatesti

Rijeka, 2016.

SVEUČILIŠTE U RIJECI

FILOZOFSKI FAKULTET

Marko Jurjako

Razlozi: naturalističko objašnjenje

DOKTORSKI RAD

Mentor: prof. dr. sc. Nenad Smokrović

Komentor: doc. dr. sc. Luca Malatesti

Rijeka, 2016.

Mentor rada: red. prof. dr. sc. Nenad Smokrović

Komentor rada: doc. dr. sc. Luca Malatesti

Doktorski rad obranjen je dana 31.10.2016. na Filozofskom fakultetu u Rijeci, pred povjerenstvom u sastavu:

1. red. prof. dr. sc. Boran Berčić

2. red. prof. dr. sc. Nenad Miščević

3. doc. dr. sc. Matej Sušnik

Acknowledgments

I would like to thank my supervisors Nenad Smokrović and Luca Malatesti for their advice and support throughout the process of writing this thesis. In addition, I would like to thank Nenad Miščević for all his help and guidance over the years. The profound impact of his philosophical ideas on my thinking is visible throughout the present dissertation.

I am especially grateful to Luca Malatesti, who gave me the opportunity to collaborate with him on the project CEASCRO (grant number 8071), which eventually resulted in my getting a two-year position as a junior researcher at the Department of Philosophy in the Faculty of Humanities and Social Sciences in Rijeka, funded by the Croatian Science Foundation (grant number 9522). This funded research position gave me invaluable time and resources to finish my thesis. Also, thanks is due to all the members of the Department of Philosophy in Rijeka who took me in as one of their own and provided me with a perfect working environment. The fact that I received this opportunity to write a PhD thesis in philosophy in Rijeka, during one of the worst economic crisis in Croatia's modern history, where the young and the old struggle to find their place under the sun, makes me even more grateful and aware of the privilege I received.

Different parts of this thesis were written in various places. Parts of the first, second, and third chapter were written during my research visit at the Konrad Lorenz Institute in Klosterneuburg (Austria) in the fall of 2014. I benefitted enormously from the interdisciplinary environment provided by the KLI. I thank the Conciliation Fund of the Republic of Austria for descendants of forced laborers or people from countries that suffered exceptionally under the Nazi regime for funding my research visit. Unfortunately, during my stay at the KLI my host advisor and the scientific director of KLI, Werner Callebaut, unexpectedly died. Werner was a great naturalist and above all a good person, sensitive to the plights of young researchers. I take this opportunity to pay my respects to his memory.

Part of the fourth and fifth chapters were written when I was a visiting researcher at Hunter College (CUNY) in New York during the fall semester of 2015/2016. My thanks go to Justin Garson, who was kind enough to support my research visit in New York, and devoted his time to mentoring me during that time. Thanks also to the whole Department of Philosophy at Hunter's for being so open about external research visitors. During my stay in New York, I had the opportunity to sit in on a class given by Sharon Street and Derek Parfit. The class centered

on Parfit's unpublished third volume of *On What Matters*. Given the topic of my dissertation, it should be clear that taking this class has proven to be priceless for me. My research visit in New York was made possible by generous funding from the Croatian Science Foundation, Faculty of Humanities and Social Sciences in Rijeka, and the Identity project (financed by the University of Rijeka). In particular, I thank the former Dean of our Faculty Predrag Šustar and the current Dean Ines Srdoč-Konestra, as well as Boran Berčić, the leader of the Identity project, for being so determined to support the careers of young researchers.

Ovom prilikom želio bih istaknuti da se zapravo veliki dio svega što sam smislio i napisao, još od doba pisanja diplomskog rada, odvilo u lošinjskom BIAS Think Labu u Nerezinama. Neizmjerno sam zahvalan "direktorima" Ivu i Sonji Brzović što su mi kroz proteklu deceniju omogućili mnoge ljetne i zimske radosti. Najveće zahvale idu mojim roditeljima, Luciu i Ivanki te bratu Ivanu koji su u svakom pogledu uvijek podržavali moje životne odabire. Oni su mi omogućili da se razvijem kao autonomno biće te da tumaranje po bespućima hrvatske zbiljnosti ne bude toliko turobno. Konačno, sve ovo što je dosad napisano ne može se usporediti sa ljubavi i podrškom koju sam kroz skoro pola svog životnog vijeka dobivao od Zdenke, mojeg *sine qua non* uvjeta. Stoga raspravu koja slijedi, posvećujem Zdenki, bez koje svijeta ne bi bilo.

Abstract

This thesis has two aims. The first one is to discuss the nature of normative reasons and to investigate which account of them would be compatible with a broadly naturalistic world view. The second aim is to show how a naturalistically constrained account of normative reasons and rationality can be fruitfully applied to some practical contexts that involve interfacing normative constraints and empirical data.

The structure of the thesis is the following: in the first chapter, I introduce the concept of a normative reason. Following the literature, I distinguish between object-based and subject-based theories of normative reasons and discuss their attractions and disadvantages. In the second chapter, I defend one type of subject-based theory, the response-dependentist view of normative reasons.

In the third chapter, I argue that subject-based theories of reasons receive support from evolutionary and naturalistic considerations. Moreover, I argue, on the basis of evolutionary considerations, that the object-based theories of reasons face serious difficulties, and therefore that we should adopt an attitude-dependent conception of normative reasons.

In the fourth chapter, I further develop a positive account of one type of subject-based theory of normative reasons. I develop a naturalistically based account of reasons that is able to account for an important distinction between hypothetical and categorical reasons.

Finally, in the fifth chapter I apply the developed framework to the case of psychopathy in order to discuss the question whether some recent neuropsychological studies show that psychopaths are irrational in their decision-making processes and behavior. I argue that current neuropsychological data do not warrant the conclusion that psychopaths are on average more irrational than other non-psychopathic individuals.

Key words: Normative reasons, hypothetical and categorical reasons, rationality, evolutionary debunking arguments in metaethics, psychopathy and irrationality

Prošireni sažetak

Normativnost se pojavljuje kao neophodan uvjet u našim promišljanjima i djelovanju. Normativni pojmovi ne obilježavaju samo našu svakodnevicu, oni su isto tako sveprisutni u filozofiji, humanističkim te društvenim znanostima općenito. Kako bi govor o normativnosti imao smisla obično se pretpostavlja postojanje nekih standarda, normi ili općenitije činjenica o *trebanju* pomoću kojih mjerimo i ocjenjujemo ispravnost vjerovanja, ponašanja i emocionalnih reakcija ili čak druge normativne standarde.

Neki autori tvrde da je relevantan smisao rečenice da *nešto treba biti slučaj* ili da *se nešto treba vjerovati* onaj u kojem tvrdnje o trebanju povlače iskaz da postoji *odlučujući razlog* (eng. decisive reason) da se to nešto učini ili vjeruje. Drugi autori tvrde da se bazični etički pojmovi mogu objasniti u terminima principa koje bi razložne osobe imale *razloga* prihvatiti ili odbaciti. Nadalje, neki autori tvrde da, pozivajući se na razloge, možemo objasniti na koji način naša volja može biti slobodna te pružiti plauzibilno objašnjenje pojma moralne odgovornosti. Dakle, jasno je da normativni razlozi zauzimaju vrlo značajno mjesto u suvremenim raspravama u etici, metaetici, političkoj filozofiji te filozofiji društvenih znanosti. S obzirom na posebnu ulogu koju bi pojam razloga trebao igrati u našim filozofskim i svakodnevnim razmišljanjima, naš je filozofski zadatak pružiti zadovoljavajuće objašnjenje tog pojma.

U slučaju normativnih razloga, problem je još istaknutiji s obzirom na to koju fundamentalnu ulogu bi trebali imati normativni razlozi. Nadalje, neki utjecajni normativisti tvrde da kako bi nešto bio razlog, mora postojati neka činjenica koja ima svojstvo *ići u prilog* (eng. count in favor of) te stvari za koju postoji afirmativni razlog. Isto tako se tvrdi da se svojstvo *ići u prilog nečemu* ne može svesti na nijednu drugu činjenicu (barem ne na ne-normativne činjenice) ili objasniti u naturalističkim terminima, tj. u terminima koji se koriste u znanostima, poput biologije, psihologije ili u kognitivnim znanostima općenito. Međutim, ako su razlozi fundamentalni za naše normativno promišljanje tada bi bila poželjna ona teorija razloga koja bi mogla objasniti na koji se način normativni razlozi uklapaju u svijet koji spoznajemo putem različitih prirodnih i društvenih znanosti.

Pisanje ove disertacije ima dva cilja. Prvi i osnovni cilj je raspraviti na koji način bi se mogla razviti teorija normativnih razloga koja će biti kompatibilna s naturalističkom slikom svijeta. Pod naturalističkom slikom svijeta mislim na objašnjenja prirodnog svijeta koje možemo pronaći u trenutno prihvaćenim znanstvenim teorijama. Drugi cilj ove disertacije je

pokazati kako se određena, naturalistički omeđena teorija normativnih razloga može na zanimljiv način primijeniti u praktičnim kontekstima.

Struktura disertacije je sljedeća. U prvom poglavlju uvodim pojam normativnog razloga te ga definiram u kontrastu s pojmom motivacijskog razloga. Razlikujem dvije teorije razloga; teorije normativnih razloga koje su usmjerene na predmet i one koje su usmjerene na djelatnika (eng. *object and subject based theories of normative reasons*). U tom kontekstu, raspravljam o tome koje su pozitivne i negativne strane tih teorija.

U drugom poglavlju branim jednu vrstu teorije razloga koja je usmjerena na djelatnike. Posebice se usmjeravam na davanje odgovora na prigovor da ova vrsta teorija nema utemeljenje u svakodnevnome razmišljanju o normativnim razlozima.

U trećem poglavlju, oslanjajući se na evolucijska i naturalistička razmatranja, branim tvrdnju da teorije usmjerene na djelatnike pružaju plauzibilnije objašnjenje razloga nego teorije usmjerene na predmet. Argumentiram da se teorije razloga koje su usmjerene na predmet suočavaju s ozbiljnim poteškoćama te stoga trebamo prihvatiti neku koncepciju normativnih razloga prema kojoj su razlozi ontološki ovisni o umu ili stavovima racionalnih djelatnika.

U četvrtom poglavlju razvijam jednu vrstu teorije normativnih razloga koja je usmjerena na racionalne djelatnike. U tom pogledu razvijam teoriju normativnih razloga koja je kompatibilna s naturalističkom perspektivom te je u mogućnosti objasniti važnu razliku između hipotetičkih i kategoričkih razloga.

Konačno, u petom poglavlju primjenjujem teoriju razloga i racionalnosti koju sam razvio u prijašnjim poglavljima na slučaj psihopatije kako bih razmotrio pokazuju li trenutno dostupna neuropsihološka istraživanja da je proces odlučivanja kod psihopata iracionalan. U tom pogledu, argumentiram kako trenutačno dostupni neuropsihološki podaci ne opravdavaju zaključak da su psihopati iracionalniji od ostalih ljudi.

Ključne riječi: Normativni razlozi, hipotetički i kategorički razlozi, racionalnost, evolucijski argumenti u metaetici, naturalizam, psihopatija i iracionalnost

Contents

1. Introduction

1.1. The normativity of reasons and the ubiquity of the normative

Normativity is pervasive and seemingly inevitable in our thought and action. Onora O'Neill gives a clear and succinct statement of how normativity is important for us, and how pervasive it is in the life of a rational person:

> Normativity pervades our lives. We do not merely have beliefs: we claim that we and others ought to hold certain beliefs. We do not merely have desires: we claim that we and others not only ought to act on some of them, but not on others. We assume that what somebody believes or does may be judged reasonable or unreasonable, right or wrong, good or bad, that is answerable to standards or norms. […] We find ourselves at sea because there is huge disagreement about the source and the authority of norms on which we all constantly rely. *(O'Neill, Introduction, 1996, p. xi)*

As the quote makes clear, normativity is pervasive in our everyday life. Moreover, normative terms do not just characterize our everyday lives but are also pervasive in philosophy, the humanities more generally, and social sciences. As it is usually claimed, in order for talk of normativity (such as talk about being reasonable or unreasonable, responding to reasons, being as one ought to be, etc.) to be sensible there needs to be some standards, norms, or more generally *ought*-facts by which we measure and validate the correctness of beliefs, conduct, and emotional reactions or other normative standards.

In recent years, there has been a noticeable tendency to claim that reasons constitute the very basis of normativity. Roughly, the idea is that the concept of a normative reason can be used as a foundational concept (Parfit, 2011a; Scanlon, 1998), on top of which all other normative notions could, in some sense, be grounded (Skorupski, 2010). The notion of a normative reason is philosophically interesting and important precisely because of the weight it is supposed to carry. Nevertheless, this is not the only reason.

Some authors claim that the relevant sense of the sentence that *something ought to be the case* or that *something ought to be believed* is that in which the ought-claim entails the sentence that there is a *decisive reason* to do that thing (Parfit, 2011a). Some authors have claimed that basic moral notions can be explained in terms of principles that reasonable people have *a reason* to accept or reject (Scanlon, 1998). Others still have claimed that by invoking reasons we can explain how our will could be *free* and provide a plausible account of moral responsibility

(Fischer & Ravizza, 1998). It is therefore clear that normative reasons occupy a very distinctive place in contemporary discussions in ethics, metaethics, political philosophy, and philosophy of social sciences (Gaus, 2011).

1.2. Explaining normative reasons: The problem

Given the special role that the concept of a reason is supposed to play in our philosophical and everyday thinking, our philosophical task is to provide a satisfactory account of this notion. However, the problem of offering an explanation of the nature of normative reasons is an instance of the problem of accounting for the phenomena of normativity in general.

Furthermore, the issue of the nature of normativity in general does not only come up in philosophical disciplines, but also in other sciences in which human agents represent the basic object of investigation. The problem with normativity is that we tend to objectify it in a way that is hard to square with a broadly naturalistic picture of the world. In particular, this difficulty stems from the role that normative concepts, especially that of normative reasons, play in our practices and thought processes. Stephen Turner nicely illustrates this point. The standard view of the antinaturalist normativists[1] is that:

> The normative is a special realm of fact that validates, justifies, makes possible, and regulates normative talk, as well as rules, meanings, the symbolic and reasoning. These facts are special in that they are empirically inaccessible and not part of the ordinary stream of explanation. Yet they are necessary in the sense that if they did not exist, ordinary normative talk, including such things as claims about what a word means or what the law is, would be unjustified, nonsensical, false, or illusory. To say that something has meaning requires that there be such a thing as a meaning. To say something is a real law is to say that there is something that validates the law as real. *(Turner, 2010, pp. 1-2)*

This view, according to which every true normative claim indicates the existence of some normative fact, begs the question as to the nature of these normative facts. The questions that are naturally raised in this context are the following:

> What is the character of this normativity that is everywhere and signaled by the presence of these terms? Is it a non-natural, non-causal property of things? A force that attaches to things, such as claims, that gives them some obligatory power? Are norms part of the furniture of

[1] The term 'normativist' is taken from Turner (2010) and refers to authors who see normativity in the humanities and social sciences as a phenomenon that cannot be reduced to explanations that are common in other more fundamental sciences, such as mechanistic explanations in physics and chemistry.

> the world, a part which is merely odd in some respects, or is it an aspect of things that are otherwise normal? Or, maybe normativity is something else entirely? Is it best understood as a kind of shadow system of rules, proprieties, scoring systems, presuppositions, and so forth that stands tacitly behind our normative practices that regulates and justifies them in a way that is hidden analogue to the way that explicit rules, scoring systems, and the like regulate and justify? And if we are bound by these things, how are we bound by them? Do we bind ourselves under norms by our commitments, or in some other way? *(Turner, 2010, p. 2)*

The acute problem here is that by introducing normative facts as basic entities that play a role analogous to that played by regular non-normative facts (such as facts about masses and forces that act on objects in accordance with physical laws) in grounding non-normative language, we are on the verge of falling into the trap of postulating gaps in our world picture that are comparable to dualisms regarding the physical and the mental, and to invoking supernatural phenomena in order to explain something that is of interest. Regarding this last point, Turner (ibid.), echoing John Mackie (1977/1990), writes that "[a] danger with these questions (…) is that by answering them in the wrong way we could make normativity into something so queer that it could not be accommodated to the rest of our ideas about the natural, explainable world."

In the case of normative reasons, the problem is even more pressing, since the nature of reasons is supposed to play such a fundamental normative role. Moreover, it is claimed by some influential normativists that for something to be a reason, there has to be some fact that has the property of *counting in favor* of that thing for which there is a reason (Scanlon, 1998, p. 18). Furthermore, it is claimed that this property of *counting in favor of something* cannot be reduced to any other fact (at least not to any other non-normative fact) or explained in naturalistic terms, that is, in terms that are used in sciences such as biology, psychology, or cognitive sciences more broadly (Parfit, 2011a; 2011b). However, if reasons are so fundamental to our normative thought, then a desirable feature of an account of reasons would be for it to explain how normative reasons fit into the natural world as "revealed by science" (cf. Harman, 2000, p. 79).

1.3. Overview of the thesis

The purpose of this thesis is twofold. First, the aim is to discuss the nature of normative reasons and to see what account of them would be compatible with a broadly naturalistic world-view. By naturalistic world view I mean accounts of the natural world that are presupposed in currently accepted scientific theories. There are different strands of naturalism that one might endorse about a domain D. For instance, one could think that naturalistic considerations *demand*

that concepts in D should be reduced to more naturalistically respectable concepts of some other domain T or that some of them should be revised—or even that they should be eliminated because they do not correspond to anything in reality. However, my aim is not to give a formal naturalistic reduction of concepts referring to normative reasons. Neither, at the beginning of my inquiry, do I feel compelled to think of normative reasons as being incompatible with naturalism and thus that some version of thoroughgoing eliminativism should ensue. My position, rather, could be characterized as methodological naturalism. Here the aim is to provide a synoptic view of normative reasons and find the best way to think of them with respect to their fundamental role, which essentially involves the ability to guide agents with certain cognitive abilities and particular social and biological histories. In this respect, my aim is not to give an account of normative reasons that will preserve all of the possible platitudes we might ascribe to reasons, or to secure their fundamental normative role. Rather, I take it that methodological naturalism, like the revisionary reductivism, leaves open the possibility that in certain respects our intuitions about reasons and their ontology should be revised.

The second aim of this thesis is to show how a particular, naturalistically constrained account of normative reasons can be fruitfully applied to practical contexts. In particular, I will show how a naturalistic account of reasons and rationality might help in interfacing empirical data and normative requirements, as they are implemented in practical cases where we are trying to decide, on empirical grounds, whether a person should be considered rational or less than fully rational. More concretely, I will use psychopathy as a case study in order to show how an account of reasons in terms of rational norms developed in the present thesis can be fruitfully applied in a practical context. In particular, I will try to show how this account might yield an answer to the question of whether people with psychopathic personality traits suffer from rational impairments or whether they simply represent a group of people whose cognitive and motivational abilities underlie normal variation in rational human capacities. This discussion should contribute to the naturalistically oriented literature that relies on empirical data in order to determine when a person is rational or suffers from rational impairments. What seems to be lacking in this area of research, such as that investigating the question of whether empirical data shows that psychopaths are less than fully rational, is an explicit account of reasons and rationality that might frame the issue in a way that helps us to see the potential answers to such questions. Thus, with respect to the second aim, the discussion of reasons and rationality that I will develop in this thesis should be seen as providing a contribution that will help us fill this gap.

The structure of the thesis is the following: In the first chapter, I will introduce the concept of a normative reason and contrast it with what is commonly known as motivating reasons. During the introductory discussion, I will rely on the explication of normative reasons as things that *count in favor of* something (Parfit, 2011a; 2011b; Scanlon, 1998). The attractiveness of this explication of reasons is that it is neutral with respect to its underlying nature (Street, 2016). Relying on Parfit (2011a), I will distinguish between object-based and subject-based theories of normative reasons and discuss their benefits and disadvantages.

In the second chapter, I will defend one type of subject-based theory of normative reasons against recent objections that such theories do not have grounding in how we ordinarily think about normative reasons and therefore lack appropriate explanatory force. In order to answer these objections, I will develop an analogy between reasons and colors, to show how our intuitions could and can be revised by scientific advancement. Furthermore, I will try to show how normative reasons that constrain people's preferences and action could emerge from motivating reasons *via* an interaction between minimally rational agents.

In the third chapter, I defend the idea that subject-based theories of reasons receive support from evolutionary and naturalistic considerations. Here, I argue, on the basis of Sharon Street's (2006) evolutionary-based argument, that object-based theories of reasons face serious difficulties, and therefore that we should adopt a conception of normative reasons according to which they are mind or attitude-dependent considerations.

In the fourth chapter, I develop one type of subject-based theory of normative reasons. Here I try to develop an account of a subject-based theory of reasons that makes sense from a naturalistic point of view and that is able to account for an important distinction between hypothetical and categorical reasons. In developing this account, I touch upon many issues, such as the relation between substantive reasons, the faculty of reason, and rationality. I rely on some models from game theory in order to explain how categorical reasons could have emerged from motivational or hypothetical reasons that people already have, and the role of different types of rationality in accounting for different types of reasons.

Finally, in the fifth chapter I will apply the developed framework to the case of psychopathy in order to discuss the question of whether some recent neuropsychological studies show that psychopaths are irrational in their decision-making processes and behavior. The question of whether psychopaths are irrational is important in many ways; not least because an affirmative answer might have deep implications about how we should think about their moral and criminal responsibility, and how to think about the proper treatments for those people who fall under the category of psychopathy. In this respect, my argument will be that current neuropsychological

data do not warrant the conclusion that psychopaths should be deemed on average more irrational than other non-psychopathic individuals. However, I leave open the implications of my arguments for broader philosophical debates where psychopathy also figures as a prominent case study.

1.4. Background assumptions and methodology

The methodology I will use, which will function as a constraint and a guiding heuristic in investigating and providing an explication of the concept of reason, is so-called *methodological naturalism.* Methodological naturalism is not exactly a set of guidance rules for how to conduct a research or write a philosophical thesis; it is a more of a philosophical statement about the general relation between philosophy and science.

I regard methodological naturalism as having two components. One component is ontological or metaphysical and the other is epistemic or methodological, construed narrowly (see Papineau, 2009). The ontological component concerns the methodological maxim of grounding concepts and purported philosophical facts, in our case, facts about reasons and rationality, "in the world of facts as revealed by science" (Harman, 2000, p. 79; see also Smith, 2012). This is just the methodological counterpart of the physicalistic/naturalistic claim "that reality has no place for 'supernatural' or other 'spooky' kinds of entity" (Papineau, 2009).

Methodological naturalism, construed more narrowly, as a claim about the philosophical practice or how philosophical activity should be conducted, is a view according to which "philosophy and science [are] engaged in essentially the same enterprise, pursuing similar ends and using similar methods" (ibid.). The complement of methodological naturalism is represented by "[m]ethodological anti-naturalists [who] see philosophy as disjoint from science, with distinct ends and methods" (Papineau, 2009).

I put the emphasis on this second component of methodological naturalism because it puts greater constraints on philosophical theories and it pertains to legitimizing arguments that are closely connected to scientific practice. In this context, it is also important to highlight that one feature of methodological naturalism (construed narrowly) is the fact that it "claims some kind of general authority for the scientific method" (ibid.). I gather that this includes the claim that a default authority should be given to the outputs of the scientific method and its presuppositions. For example, this constraint would allow us to claim that legitimate norms of

rationality are those that are derived or at least underpinned by the relevant scientific practice or theories that use the concept of rationality (see Colyvan, 2009).

Furthermore, methodological naturalism (narrowly construed) forces us to ask certain questions and frame our investigations in certain terms, which will hopefully allow us to see where the theoretical problem is, what methods to apply to solve it (if it is solvable), and how feasible the solution is. In particular, it will be very useful to question the function of the concept of a reason in our discourse, and to ask what role it plays in our mental economy. In other words, methodological naturalism, as I understand it, forces us to frame the issue in terms of the problem that human beings (or rational agents more generally) are facing or trying to deal with, so that the acceptance or the introduction of the notion of a reason could help us to solve this.

The approach recommended by methodological naturalism stands in stark contrast to the traditional conceptual analysis approach of analytic philosophy (Jackson, 1998). The standard idea of conceptual analysis proceeds by proposing analyses of concepts and then testing them against our intuitions about the concept's application; if we can find a counterexample then the proposed analysis fails, if not then the analysis may be deemed successful. The whole process is performed *a* priori, without relying on judgments based on contingent experience. The most famous example of this methodology is the case of the so-called Gettier problem. In his (1963), Edmund Gettier shows that our intuitive belief that the concept of knowledge can be analyzed in terms of justified true belief is wrong, because we can imagine cases (counterexamples) in which a person has a justified true belief, but where we would not ascribe knowledge to that person.

On the other hand, methodological naturalism proposes that we do not completely rely on our *a priori* intuitions about concept application. Rather, it recommends that we rely on the way in which the relevant concepts are used in successful scientific theories. Thus, part of the task of methodological naturalism is to see how our ordinary concepts interface with scientific concepts (such as the folk-psychological concept and the scientific concept of rationality). Furthermore, this approach puts constraints on concept application that does not solely stem from our *a priori* intuitions; it will also depend on the actual usage that we observe in scientific theories.

Since methodological naturalism contrasts with traditional *a priori* conceptual analysis, I should say something about why I feel I am justified in taking the former approach seriously. The answer to this question might not be completely compelling because there is no argument

that can persuade everyone to accept naturalism.[2] Some even claim that naturalism cannot be given any completely non-circular argument in its favor (Giere, 2008). But that is as it should be, since philosophical naturalism does not aim to offer special foundations for scientific practice and thereby validate it and its role in a philosophical theorizing. Rather, philosophical naturalism sees its role in continuity with the sciences, differing from the rest of the sciences by being occupied with more abstract and conceptual issues (Quine, 1981).

However, I adduce two considerations that seem to favor the adoption of methodological naturalism as propounded here. One is the idea—or by now the platitude—that science is our most successful endeavor to explain the nature of the world and our place within it. The naturalistic hope is that staying close to science will have beneficial effects and hopefully provide new perspectives on hard philosophical issues.

This naturalistic stance is based on an inductive inference that is often used for arguing in favor of the causal closure of the physical domain, namely the ontological idea that all *physical effects* can be traced back to *physical causes* (see Appendix in Papineau, 2004). Nevertheless, in my discussion I do not rely on the principle of causal closure of the physical. I take it, rather, that the inductive inference from the past and present success of empirical sciences at least warrants paying attention to the relevant empirical sciences and trying to ground or interface philosophical concepts with explanatory concepts found in the relevant empirical theories. In recent decades, we have seen the benefits of this approach in the investigation of the evolutionary, neurological, and cognitive underpinnings of morality. Here philosophers' engagement with scientific data enabled formulations of new perspectives and arguments that advanced the debate on traditional issues such as the nature of moral judgment and its relation to motivation. In turn, these philosophical engagements enabled further formulation of scientific hypotheses and lines of inquiry. And there is currently no reason to think that rigorous scientific methods of investigation and theorizing cannot be applied to topics, such as ethics, that are still considered to be primarily philosophical, because even regarding ethics scientific probing has been underway for some time now (see e.g. Sinnott-Armstrong, 2008). In this respect John Doris and Stephen Stich write:

> The most obvious, and most compelling, motivation for our perspective is simply this: It is not possible to step far into the ethics literature without stubbing one's toe on empirical

[2] Some even claim that if there were reasons to accept naturalism then naturalism would be false, since it cannot accommodate the notion of a normative reason (Parfit 2011a). This issue will be dealt with in the following chapters.

> claims. The thought that moral philosophy can proceed unencumbered by facts seems to us an unlikely one: There are just too many places where answers to important ethical questions require—and have very often presupposed—answers to empirical questions. *(Doris & Stich, 2012, p. 112)*

These considerations bring us to the second point.

The other reason for adopting methodological naturalism stems from the pragmatic agenda to which this thesis is committed. First, one of the aims of this thesis is to see what kind of conception of reasons emerges if we hold fixed scientific knowledge and theories that are relevant for the present issue. Second, at least *prima facie*, an account of reasons and rationality would seem to be better if it can be used to interface notions of reasons and rationality as they are used in ordinary practices with their potential counterparts in the empirical sciences. For example, one interesting issue is how empirical data on human reasoning capacities can be used to determine whether people are rational or whether and to which degree they respond to reason (Samuels, Stich, & Bishop, 2002). In particular, the issue of whether psychopaths are rational immoralists has recently emerged (Aaltola, 2014; Maibom, 2005). Here, the main evidence that has been used in adjudicating the question is data from neuropsychological studies (Jurjako & Malatesti, 2016). Third, given the second point, I will take naturalistic considerations into account to develop an account of reasons and rationality that may be sensibly used to discuss some of these issues.

The fact that we need to rely on scientific data immediately puts us in a position that gives default authority to those data and the empirical theories that explain them. Furthermore, empirical theories that explain scientific data carry constraints on what kind of concept of reason we might adopt and what norms we might expect to govern the capacities denoted by the concept. For example, it is natural to think about rationality as the capacity to adaptively respond to present and future environments given one's aims and values. This conception of rationality is also used in different accounts of criminal and moral responsibility (Fischer & Ravizza, 1998). In addition, it is plausible to think about these capacities as *executive functions* that are implemented in the brain's prefrontal cortex. However, when we start to think about reason or rationality as implemented in the brain's functions then we need to be sensitive to those functions that cannot be determined *a priori*. We need to think in terms of what the brain is doing as implemented in the body and its function in regulating behavior and different processes in the body. Plus, we need to be sensitive to the brain's evolutionary history and why rationality as an executive function might have evolved. This external perspective on the functions of rationality and its implementation will, in turn, *a posteriori* constrain, *via* our

scientific theories, which norms we can legitimately think of as governing the proper operation of rationality and by extension which reasons we can ascribe to people.

In the case of people with psychopathic personality disorder, we need to be sensitive to these issues of interfacing since there is more and more pressure to determine the status of their rationality and responsibility through hard data from neuropsychological studies (Focquaert, Glenn, & Raine, 2015; Sifferd & Hirstein, 2013). These considerations force us to take methodological naturalism seriously, according to which *a posteriori* presuppositions coming from the sciences should constrain and direct our arguments and the formation of our theories. However, in committing myself to this methodological conception I still recognize the importance of the concepts with which we started our investigation. In this sense, I endorse José Bermúdez's warning that:

> we must not forget that the obligation of answerability goes in two directions. Our scientific investigations must be sensitive to our pre-theoretical understanding of the concepts in question, but so too must we be prepared to change our pre-theoretical understanding in response to what we learn from empirical investigation. *(Bermúdez, 2005, pp. 12-13)*

2 Accounts of normative reasons

2.1 Introduction

The purpose of this chapter is to explain what normative reasons are and lay out the main views regarding their nature. The main aim is to present two families of theories that try to account for the sources of normative reasons. For the sake of clarity regarding the main topic of this chapter, as well as the problem I will try to solve, I will first delineate different senses that the word 'reason' might have in normal contexts, and then isolate that which is most relevant to the present discussion, namely the concept of a normative reason.

In the first part of this chapter (sections 2.2–2.3), I will delineate the notion of a normative reason and discuss its relevant structural features. Then, in the second part (sections 2.4–2.5), following Derek Parfit (2011a) I will distinguish between two accounts of normative reasons, object- or value-based theories and subject-based theories of normative reasons. The aim of this chapter will be to show, first, the main problem with object-based theories, and second, to highlight the problems facing subject-based accounts and to show how they could be overcome. Regarding the last point, I will present the main challenges that Parfit puts to subject-based theories, and try to show how a subject-based theorist might plausibly answer them.

2.2 What are reasons?

2.2.1 Two general kinds of reasons

The concept of reason plays multiple roles in normal conversational contexts. In certain contexts, the term 'reason' is synonymous with the term 'cause'. For example, we say that the reason why a building collapsed is the fact that an earthquake occurred. In the case of human action, we also use the concept of reason to explain why someone did something. For example, we might wonder why Smith robbed a bank. The answer might be that he wanted to get some extra money so that he could pay for some very expensive medical treatment for his sick grandmother, and that he believed that by robbing the bank he would be able to afford the

treatment for his grandmother. In this example, the desire to help his grandmother and the belief about the likely means of doing so represent the *reason* why Smith robbed the bank.[3]

Similarly, the concept of reason can be used to explain the formation of mental states, not only observable behavior. Thus, we can explain why Smith believes that his grandmother is very sick by providing a reason explaining the formation of his belief. In our imagined example, the reason why Smith believes that his grandmother is sick could be the fact that his fortune-teller told him so. Furthermore, we can imagine that she has told him that if he does not act promptly his grandmother will soon die.

The reasons I have mentioned so far are standardly called explanatory reasons because their role is to explain why something happened or to indicate what the cause of some event was. In the sphere of practical philosophy, explanatory reasons are usually called motivational reasons, because they explain the actions of an agent by citing a motive for which the agent acted (see e.g. Lenman, 2009). Thus, explanatory reasons explain the *factive* dimension of reality, that is, they explain why things are *thus-and-so* or why something happened or is happening, etc.

Explanatory reasons are utilized in predicting and explaining behavior and formation of the mental states of individual agents. It is standardly assumed, following the Humean philosophical tradition (see e.g. Davidson, 2001, essay 1; Smith, 1987), that explanatory reasons are composed of a pair of mental state-types, composed of beliefs and desires. The theory that utilizes concepts of desires and beliefs in order to explain and predict agential behavior is, in philosophical circles, often referred to as *folk psychology*, and in cognitive science literature as *Theory of Mind* (Ravenscroft, 2010).[4]

Explanatory reasons are contrasted with *normative* or *justificatory* reasons (Lenman, 2009). Generally, one can say that normative reasons indicate how things *should* or *ought* to

[3] The explanatory scheme that utilizes the notions of desire and belief in accounting for behavior or intentional action is called folk psychology. Generally, when we use the latter to ascribe mental states (such as beliefs and desires) to other organisms or persons this is standardly called *the theory of mind*.

[4] The use of folk psychology or theory of mind for explaining and predicting behavior or mental states is called *mindreading* (Ravenscroft, 2010). Mindreading usually proceeds by attributing mental states to a subject, and then on the basis of those mental states a prediction or an explanation of the subjects' action or formation of other mental states is extracted. For example, if the action has already been performed, we can explain Smith's behavior by saying that he wanted to get some money in order to be able to pay for proper treatment for his grandmother and that he believed that by robbing the bank he could effectively achieve this goal. The ability to mindread starts to develop in infancy and it seems to mature in children at the age of 4 (Wimmer & Perner, 1983). The evolutionary origins of the theory of mind are still debated and whether or not the capacity for mindreading should be attributed to non-human primates is still a matter of controversy (Call & Tomasello, 2008).

be, and not how things really are or will be in the future (or how we predict them to be). It could also be said that normative or justificatory reasons play the role of instructions, that is, they indicate the desirability or worthiness of states of affairs and thus, are often connected to what is considered to be valuable (Parfit, 2011a, pp. 38-39; Raz, 1999, ch. 2).

This special feature of normative reasons is usually unpacked by saying that "reasons are considerations that count in favor of that thing for which they are reasons for" (Schroeder, 2007, p. 11; compare Parfit, 2011a, p. 31; Scanlon, 1998, p. 17). With this terminology in mind, one can often hear that certain facts *favor* the adoption of certain attitude towards some proposition. For example, there is a proposition that the high concentration of iridium at the cretaceous-tertiary boundary *counts in favor* of the thesis that in that period of Earth's history an asteroid fell on Earth, which caused the extinction of dinosaurs. In the practical domain we often encounter statements such as the following: the fact that smoking cigarettes is bad for your health *counts in favor of* stopping smoking, or the fact that some group of people will benefit from acting in a cooperative manner counts in favor of acting or being moral (see e.g. Gauthier, 1986).

From the point of view of normative reasons, we can take another look at the Smith example and see where the difference between normative and explanatory reasons is most salient. In the example, Smith formed the belief that his grandmother was not healthy because Smith's fortune-teller told him that this was the case. However, from the normative perspective we can criticize Smith's formation of the belief that his grandmother is sick because fortune-tellers are not very reliable sources of information and therefore do not provide good reasons for believing what they say. Furthermore, we can criticize Smith for robbing the bank because he did not have a very good reason for doing it, because, for example, from a moral point of view it is wrong to steal and induce unnecessary pain in other people. This can be the case even though we can recognize and understand Smith's reasons for performing the action.

The general point is that one can have a reason for believing or doing something without that reason being good in the normative sense, that is, without it counting in favor of that thing. On the other hand, one can have a normative reason for doing something without having a motivational reason for doing that thing. An explanation for this situation could be that the agent does not recognize the reason or that the agent recognizes the reason but simply does not respond to it.

As an illustration of the first case, we can take the famous example given by Bernard Williams (1981). In the example we have a person who comes into a bar and orders a gin and tonic. Unbeknownst to her, the bartender pours petrol into her glass. However, since she thinks

that she has been given a gin and tonic she takes the glass and drinks the petrol. Her desire to drink gin and tonic and her belief that the glass in front of her contains gin and tonic give us an explanation (a motivational reason) for why she drank from the glass. Nevertheless, intuitively, this example shows that even though she had a motivational reason to drink the contents of the glass she did not actually have a *normative reason* to drink it. That is, we can say that the fact that the glass contained petrol counted against her drinking from the glass.

As for the second case, the standard example in the literature is the phenomenon of *akrasia,* or weakness of the will. As is often the case, a person knows that smoking cigarettes causes cancer and that this fact counts strongly in favor of stop smoking. Nevertheless, the person, because of weakness of will, continues to smoke when the opportunity arises even though she recognizes that it would be better if she stopped smoking.

In this chapter, and in the rest of the thesis, the emphasis will be on the normative reasons or on the normative *dimension* of reasons.[5] In developing this topic, I will rely heavily on Derek Parfit's recent influential work (Parfit, 2011a; 2011b). There are at least two reasons for my choice. One is that Parfit wrote two volumes dedicated to the discussion of normative issues, where reasons play a special and central role. Thus, he provides a framework for discussion about normative reasons. The other is that Parfit exposes problems that the concept of a normative reason introduces into the naturalistic picture of the world. In arguing for his brand of normative realism, Parfit provides intuitively strong arguments against naturalism about normative reasons. Hence, besides providing a framework for talking about reasons, Parfit serves as an opponent whose arguments should be disarmed in order to show that normative reasons can be incorporated into a naturalistic picture of the world.

2.3 Normative reasons

Among normative reasons there is a standard distinction between theoretical or epistemic reasons and practical reasons or reasons for action. Broadly speaking, the distinction between theoretical and practical reasons can be provided in folk-psychological terms. Using folk psychology we can explain and predict the behavior of intentional agents by attributing them

[5] I write normative dimension of reasons because the explanatory and normative reasons from which or in accordance with which an agent acts or forms a belief can often be the same. For example, we can say that the reason Smith believes his grandmother is sick is that a qualified doctor has examined her and gave Smith the diagnosis. In this case, the fact that the doctor told Smith that his grandmother is sick is a reason that explains Smith's belief and gives a justification for his belief.

mental states. The attributed mental states are standardly divided in two broad categories, which can roughly be termed cognitive states and motivational states. Among the cognitive states we find beliefs, suppositions, assumptions, plausibility judgments, etc. Among the motivational states we have desires, intentions, emotions, preferences, and so on. For ease of discussion, cognitive states are often lumped together under the term 'belief' and motivational states are subsumed under the term 'desire' (Smith, 1987). Because of this terminology, folk psychology is often called belief–desire psychology.

Using the latter classification, the distinction between epistemic and practical reasons can be made in terms of reasons that support different kinds of mental states. Thus, theoretical reasons are reasons to believe something or to adopt a belief about a certain state of affairs. For example, the fact that scientists found iridium in the cretaceous-tertiary part of the Earth's crust provided them with a reason, given their other theoretical beliefs, to believe that dinosaurs died out due to the impact of an asteroid (for more on epistemic reasons see introduction in Reisner & Steglich-Petersen, 2011). In contrast to epistemic reasons, practical reasons count in favor of actions, desires, and intentions; more generally we can say that they are about what kind of motivation one *should* have. For instance, we normally take it to be the case that if a person is in pain we have a reason to help her alleviate that pain.

However, the broad distinction between epistemic and practical reasons is probably more intuitively understood in terms of rational requirements that apply to motivational and epistemic states. Gilbert Harman (2004) gives a good example of different requirements that apply to intentions and beliefs. For example, let us say that I am trying to decide on the best way to get to my place of work and I realize that there are at least two optimal routes. That is, taking either of them demands a similar amount of effort, they are of equal distance, they are similarly boring, similarly safe, etc. These features of the routes make it rational for me to choose one arbitrarily. Since the two routes are similar in all relevant respects, it is completely rational to choose which one to take by flipping a coin, for example. However, in the epistemic case the analogous situation would not warrant the arbitrary adoption of a belief. For example, if I am in a situation in which I have equally strong evidence that p is the case and that not-p is the case, then, epistemically speaking, I am not allowed to arbitrarily adopt the belief that p is the case or the belief that not-p is the case. Rather, the epistemically rational response would be to suspend judgment.

These considerations about the rationality of forming different attitudes enable us to see the difference between practical and epistemic reasons. Practical reasons seem to be considerations that satisfy the rational requirements that apply to practical attitudes, such as

intentions, in our example. Epistemic reasons are different; they are considerations that satisfy the rational requirements that apply to beliefs, for example. Thus, we see that intuitively there is a tight connection between the facts that represent reasons of different types and rational requirements that apply to the different attitudes for which we seek reasons.

Besides the broad division between epistemic and practical reasons, we can also talk about further subdivisions between normative reasons. For example, we can think about aesthetic reasons, reasons of etiquette, moral reasons, legal reasons, and so on. In this context, we might ask about the relation between, for example, epistemic and practical reasons, and whether one kind can be reduced to the other. However, for our present purposes this issue is not important. In what follows I will say more about the general concept of a normative reason in order to pinpoint some structural features that will then provide a platform for further discussion about the nature of normative reasons.

2.3.1 Platitudes about normative reasons

It is standardly assumed that reasons have certain features that can be read off from the following general form of the reason-relation provided by John Skorupski (cf. Skorupski, 2010, p. 37):

Set of facts r_i *is at time t a reason of degree of strength d for X to* ψ.

Where r_i stands for facts that count in favor of something (the ground or the basis of the reason relation), t stands for time, d for the strength of the reason or reasons in question, X for an agent to whom the reason relation applies, and ψ for the thing the grounds are reasons *for*, whether it is a belief, action, desire, or some other attitude.[6] Usually in discussions about normative reasons reference to time is omitted, so in my discussion I will also not say much about the temporal dimension of reasons.

To give an intuitive example, we can say that the fact that this glass contains petrol is a strong reason now for Mary *not* to drink from it. Alternatively, we can say that the fact that Smith has seen the fossil records of different organisms is a reason for him to believe that evolution occurred.

[6] Skorupski (2010, pp. 35-36) thinks there are three basic types of reason-relations: reasons for belief, action, and feeling, which he terms epistemic, practical, and evaluative reasons, respectively.

2.3.2 Reason is a relation

The formal structure of propositions that employ the concept of reasons shows that reason is a *relation* between facts and attitudes.[7] It also shows that the relata of the reason-relation include a basis or a ground that is constituted by some facts and an attitude for which those grounds count in favor of the attitude. So, the fact that the clouds are grey, for example, is a ground of the reason-relation that supports the belief that it will rain, which is the other relata of the reason-relation. And of course the fact that the clouds are grey supports the belief to a certain degree, because the relation between the fact that the clouds are grey and the fact that it is going to rain is only probabilistic.

It is important to stress that reason claims are relational because often reasons are identified with facts that constitute the ground or the basis of the reason-relation. There is not much harm in doing so, but in stressing that reasons are relations that are captured by the phrase 'counts in favor of' we can avoid some possible conundrums, such as how reasons can be normal descriptive facts, like the fact that I am in pain, and at the same time be normative in the sense of indicating that something needs to be done. The solution is to see that the fact that I am in pain cannot be identified with the fact that I have a reason to change my situation. It is better to say that the fact that I am in pain *counts in favor of* changing the current situation in some way.[8]

2.3.3 Pro tanto and prima facie reasons

The degrees of support that reasons bring with them indicate their *pro tanto*[9] nature (Broome, 2004). *Pro tanto* reasons are those reasons that genuinely count in favor of ψ-ing, but it might

[7] Earlier I said that practical reasons could be reasons for action; however actions are not attitudes. To bridge this apparent gap we can say that reasons for action are mediated by reasons for intention. And since intentional action normally springs from an intention to perform an action, we can preserve the connection between reasons and attitudes.

[8] For more on this see Skorupski (2010).

[9] Different authors express the idea that reasons can be *pro tanto* differently. For example, they used to be called *prima facie* after the distinction made by David Ross (1930) between *prima facie* and absolute duties. However, the term '*prima facie*' implies that what we thought was a reason could turn out not to have been a reason in the first place. To use Williams' example again, the fact that Mary ordered gin and tonic is a *prima facie* reason to drink the stuff in the glass that was given to her by the bartender. But the fact that the glass contains petrol cancels out the *prima facie* reason. That is, were Mary to find out that what is in the glass is actually petrol, she would realize that she actually does not have a reason to drink the stuff in the glass. The term '*pro tanto*' allows even outweighed reasons to retain their status as reasons that count in favor of something. For the example of *pro tanto*

be the case that their degree of support for ψ-ing is not decisive, that is, it could be outweighed by other, stronger reasons (cf. Lenman, 2009). For example, the fact that the glass contains petrol is a reason not to drink from it. However, it might be the case that Mary's drinking petrol will save a person's life. We can suppose that some malicious people threaten to kill Mary's friend unless she drinks the petrol from the glass. In that case the fact that drinking the petrol could save a life could be a reason for Mary to drink the petrol or at least to form an intention to drink it. However, the fact that drinking petrol could make Mary sick is still a reason not to drink it, albeit a reason that is outweighed by the stronger reason to do it and save a friend's life. Lenman (2009) provides another example. We can suppose that the fact that smoking gives Mary pleasure provides a *pro tanto* reason for her to smoke. However, even though we might think that there is something speaking in favor of Mary's smoking, we might nevertheless think that *all things considered* Mary should not smoke.

Reasons can also be *prima facie*. Unlike *pro tanto* reasons, *prima facie* reasons can be defeated and not just outweighed. Let me illustrate the point by using Williams' petrol example again. The fact that Mary ordered a gin and tonic is a *prima facie* reason for her to drink the stuff in the glass that was given to her by the bartender. But the fact that the glass contains petrol cancels out or defeats the *prima facie* reason. In other words, if Mary were to find out that the glass actually contains petrol, she would realize that she in fact does not have a reason to drink the stuff in the glass. Thus, when Mary realizes that there is a *defeater* of her reason to drink the stuff in the glass, that reason *stops* counting in favor of drinking the stuff in the glass.

John Pollock (1987, p. 485) distinguishes between rebutting and undercutting defeaters. Rebutting defeaters are those that defeat a *prima facie* reason by contradicting the conclusion of the reason-relation. The petrol example illustrates the rebutting defeater. The fact that Mary ordered gin and tonic counts in favor of drinking what is in the glass she is given by the bartender, but the fact that the glass contains petrol counts in favor of not drinking the stuff in the glass and because of this, the latter reason defeats the former.

Undercutting defeaters are those that undermine the connection between the reason and what the reason counts in favor of (the conclusion). Thus, we can say that the fact that it looks to us like Smith is in pain is a reason to help him. Nevertheless, the fact that we are in a theater and Smith is an actor undercuts the conclusion that we should help him. Although the fact that

reasons see the main text. Dancy (2004) uses the expression 'contributory reason' for what is nowadays usually called a *pro tanto* reason.

Smith is playing in a theater undercuts the reason to help him, it still does not mean that we have an opposite reason not to help him, since, as the case may be, he really might be in pain.

2.3.4 Reasons and deliberation

One of the most important roles that reasons play in our thinking is in deliberation about what to do or what to believe (Enoch, 2011, ch. 3). When choosing between different possible acts or deciding what to believe, reasons that bear on the issue can conflict. Intuitively speaking, the rational choice of an action or endorsement of a belief depends on the strength or weight, or we might even say the force of reasons (cf. Parfit, 2011a, p. 32). Reasons may be combined so that, on the one hand, we have a strong reason to do one thing, but on the other hand, we have several, separately weaker reasons, that when combined become stronger than the first. Parfit gives an intuitive example:

> If I could either save you from ten hours of pain, or do something else that would both save you from nine hours of pain and save someone else from eight hours of pain, I would have a stronger set of reasons to act in this second way. As we can more briefly say, I would have more reason to act in this way. *(cf. Parfit, 2011a, p. 32)*

Parfit also introduces the concepts of *decisive* and *sufficient* reasons. We have a decisive reason to act when "our reasons to act in some way are stronger than our reasons to act in any of the other possible ways." Furthermore, Parfit tells us that acting in accordance with the decisive reason "is what we have *most reason* to do" (ibid.).

However, the concept of a sufficient reason is introduced because intuitively there will be some situations in which there will be no decisive reason to do any particular thing, but still *enough* reason to act in more than one way. Thus, Parfit writes, "our reasons are sufficient when these reasons are not weaker than, or outweighed by, our reasons to act in any of the other possible ways" (cf. Parfit, 2011a, p. 33). To illustrate the point, he gives an example:

> We might have sufficient reasons, for example, to eat either a peach or a plum or a pear, to choose either law or medicine as a career, or to give part of our income either to Oxfam or to some other similar aid agency, such as Medecins Sans Frontieres. *(cf. Parfit, 2011a, p. 33)*

Reasons and their role in deliberation can be of different levels. We might have a first-order reason do to ϕ, but also a second-order reason to disregard the first-order reason in a particular situation.[10] Joseph Raz (1975, p. 34) defines second-order reasons as "reason[s] to act on or

[10] The terminology of first- and second-order reasons comes from Joseph Raz's influential (1975) book.

refrain from acting on a reason."

Thomas Scanlon (1998, p. 51) gives the following example that can be used to illustrate the distinction: when we play tennis we have to decide whether we are going to play competitively or not. Let us say that we decided to play competitively. In that case, the fact that a certain shot represents the best strategy for winning a point will be a sufficient reason to perform it. With this in mind, we will not have to weigh the reason with the possibility that our opponent will in some way suffer or that we will hurt her feelings as a result of playing competitively. Even though it might be the case that we have reasons to care about our opponent's feelings, they are not going to be taken as relevant in a situation in which the two of us play tennis. Thus, in this example, we could have a first-order reason to care about the feelings of our opponent. Nevertheless, during the tennis match those reasons would be disregarded because of our second-order reasons, which in this case involve the decision to play competitively.

The normative element of reasons comes to the fore when we introduce the concept of an *ought* into talk about reasons as facts that count in favor of something. That is, the connection between what ought to be the case and the favoring relation comes to the fore when we think about what to do or believe, and then reach a judgment about what we have a decisive reason to do or believe. In that case, it is natural to say that what we have a decisive reason to do is what we should or ought to do. According to Parfit (2011a) the general sense of ought that is important in the context of normative reasons is that which implies that there is a decisive-reason counting in favor of what ought to be done.

This makes intuitive sense, because when we ask why I should Φ or believe that p we are asking for a reason, and that reason should in some sense explain[11] or justify the ought-claim. Thus, someone could tell John, who is a wealthy person, that he ought to help Smith by giving him some money. John could then ask why he should help Smith by giving him his hard-earned money? In this situation, one could say to John that Smith is his friend and that Smith does not have enough money to provide treatment for his sick grandmother, and moreover, that John has more than enough money to take care of himself even if he helps Smith. After John is provided with reasons that, supposedly from his point of view, justify the claim that he should help Smith, he can reach a decision on the basis of the fact that all relevant considerations count in favor of the claim that he should help Smith. In other words, John can reach a decision that all things

[11] See Broome (2004; 2013) for a development of a reductive account of reasons according to which reasons are facts that explain why something ought to be the case.

considered he has a decisive or at least sufficient reason to help Smith.

Other forms of deliberation can take place in private thought, such as when one tries to decide what one has a reason to do the following weekend. For example, one can deliberate about whether to visit a Zoo where they have a new and exotic animal or whether to visit a gallery where Picasso paintings are exhibited. For both options there are presumably some reasons that could be adduced in their favor, and the role of deliberation is to weigh and balance those reasons in order to reach a conclusion about what one has most or at least sufficient reason to do. Therefore, we can add that reasons also play a role in determining what one ought to do in this deliberative sense.

2.3.5 Reasons, rationality, and advice

Other platitudes about reasons include the idea that reasons are those things that could be offered as advice about what to do or what to believe (Smith, 1994; 2004). The fact that the glass contains petrol is a reason for Mary not to drink from it, and because it is a reason for Mary not to drink it, a person that is in a better epistemic position than Mary could offer this fact as a piece of *advice* to Mary not to drink from the glass.

This idea is related to the fact that an agent does not have to be aware of all the normative reasons that apply to her in a particular situation. On the basis of this connection between advice and reasons, Michael Smith (1994; 2004) has developed an account of normative reasons according to which an agent has a reason to Φ if her rational self would desire that she Φ, that is, if her rational counterpart, who knows all the relevant information about her and her circumstances, would *advise* her to Φ.

We can make another distinction in terms of the relation between rational advice-giving and awareness of reasons. The distinction is between subjective and objective reasons, that is, between reasons that an agent *believes* she has and those that really apply to her. When one is not aware of a reason then one might do what one *ought not* to do from the perspective of that reason. A person can, so to speak, act against that reason. However, even if that is the case, the person in question can still be rational given the beliefs in the light of which she acts. The petrol example illustrates this point well. Mary has a reason not to drink the stuff in the glass, however, if she actually drinks the petrol, she will still be perfectly rational, at least in a minimal sense. Because she would be acting according to her justified belief that what is in the glass is what she ordered. In a derivative sense, her action would be justified from her own subjective perspective. That is, she would be acting for a reason that she *believes* obtains.

Thus, we see that the notion of rationality, which could be related to a more subjective notion of a reason, has this more intensional dimension. According to this subjective conception, what is rational to do or believe largely depends on what a person wants and believes. This conception is opposed to the notion of an objective reason that is more extensional in the sense that what there is a reason to want and believe will depend on facts and not strictly on the mental state of the agent. Once more, we can use Parfit's example to illustrate the point:

> Suppose that, while walking in some desert, you have disturbed and angered a poisonous snake. You believe that, to save your life, you must run away. In fact you must stand still, since this snake will attack only moving targets. Given your false belief, it would be irrational for you to stand still. You ought rationally to run away. But that is not what you ought to do in the decisive-reason-implying sense. You have no reason to run away, and a decisive reason not to run away. You ought to stand still, since that is your only way to save your life. *(Parfit, 2011a, p. 34)*

The example illustrates that what we think is rational can diverge from what we, from a third-person point of view, think there is a reason to do and would advise ourselves to do. The gin-and-tonic example and the disturbed-snake example are supposed to show that our intuitions about what is rational and what we have reason to do can part ways. From these considerations, some authors conclude that the intuitive idea according to which rationality consists in responding to reasons must be false or at least not so straightforward, since one can be rational even though one does not respond to an externally determined reason (Broome, 2013). According to this view, rationality could be construed as a set of requirements that put constraints on the appropriate combination of mental states, whatever reasons there are for holding those attitudes. For example, on this view rationality would require you to intend to run away given that you want to stay alive and believe that you will do so only if you run away. If you fail to form the intention to run away given your other mental states, it would seem that you end up having an irrational combination of mental states, no matter what the external facts are. On this view, reasons and rationality could part ways, because reason would require that you stay put (unbeknownst to you), while rationality would require that you run away, given your present attitudes.[12] Thus, some authors contend that rationality is different from reasons,

[12] Actually, for proponents of the view that rational requirements are wide in scope, rationality would require one either to form an intention to run away, or to change one of one's attitudes in order to restore coherence between one's mental states. This could be intuitively illustrated with theoretical reasoning that involves *modus ponens*. If we agree that modus ponens inferences present one of the requirements of rationality, then rationality according

because reasons depend on external facts while rationality supervenes on mental states, regardless of what facts are reasons for what (Broome, 2013).

However, introducing a conceptual separation between rationality and reasons seems to lead to further questions. We normally think about rationality as being normative in some sense, that is, as being such that we ought to follow the rules of rationality and that there is something wrong if we break them. However, if rationality consists solely in satisfying some coherence criteria on the admissible formation of beliefs and does not have anything in particular to do with responding to reasons, then we can ask ourselves why we should be rational or what reason we have for being rational (Kolodny, 2005). If we think that rationality simply demands coherence among our attitudes, then it is hard to see how we could give a principled answer to the latter question and to say what would be wrong with failing to be rational.

Nevertheless, examples in which judgments of rationality and reasons go separate ways do not necessarily break the intuitive conceptual connection between reasons and rationality. As we saw above, it is natural to think about reasons as pieces of advice that someone in a better epistemic position could give us. So, naturally we can extend that idea by saying that reasons are those facts which, if we were fully rational, we would use to give ourselves advice about what to do or what to believe. In the angry-snake or gin-and-tonic examples, we can say that we are not fully rational because we lack an important true belief, and thus our rational capacities fail to track what we really have a reason to do. However, failing to be fully rational does not necessarily mean that we are in a culpable state, especially not when the circumstances are unusual. On this account, the question of why I should be rational is closed, at least if by this question we ask what counts in favor of being rational. Since being a fact that counts in favor of something is just being a fact to which rational agents respond, the question reduces to asking what counts in favor of my responding to facts that count in favor of doing something.

---

to wide-scopers would require something of the following sort: let us say that agent A believes that p, believes that if p then q, and believes not-q. Since this combination of beliefs is inconsistent, rationality requires that either that A stops believing that not-q or that she revises the belief that p, since revising one belief or the other would restore consistency between them. Formally, *modus ponens* as rational requirement could be spelled out in the following way by using an ought operator that has a wide scope: Ought(if believe that p $\wedge$ Believe that (p $\rightarrow$ q)), then believe that q). From this requirement we cannot conclude that one ought to believe that q given the other beliefs. Rather, one satisfies the requirement either by not believing that the conjunction (p $\wedge$ (p $\rightarrow$ q)) is true or by simply believing that q is true. See Broome (2013).

Here we seem to hit bedrock, because if there is a reason to do something then it seems that that reason reflectively provides a reason to respond to it.[13]

Other authors drop the reference to *full* rationality and explaining reasons in terms of rationality and simply say that rationality consists in responding to *apparent* or *subjective* reasons (Parfit, 2011a; Schroeder, 2007). Here apparent or subjective reasons represent those considerations that *would* be objective reasons if our relevant beliefs were true. For instance, in the angry-snake example, it is rational to run away because I would have a decisive reason to run away if the belief that by running away I would save my life were true.

To generalize these ideas, we can say that the function of rationality is to track reasons. In particular, rationality tracks reasons when background conditions are normal. We would minimally include having relevant true beliefs among background conditions. Thus, if background conditions are normal, exercising our rational capacities will tend to lead us to what there are reasons to do. However, if background conditions are not normal then either rationality could mislead us, such as when we act on the basis of a false belief, or there might be a defect in rationality that would lead to irrational behavior, such as when we act in ways that are self-defeating (e.g. we run away despite knowing that we should stay put).

2.3.6 Overview: The structure of reasons

What has been so far said about the structural features of reasons can be summed up in **Table 1**. See the next page.

[13] This principle might be called the *iterativity of reasons*, which says that among the reasons that we have, there is also a reason to respond to reasons (cf. Johnston, 1989, p. 158).

Table 1

1. Reason is a relation between facts and attitudes and so it has directionality. Reason can count *for* or *against* having an attitude or performing a certain action.
2. Reason has a basis or ground constituted by some facts or propositions.
3. What is a reason *for* is usually taken to be some kind of attitude. The *for* part indicates the direction of the reason.
4. Bases or grounds have strength. In other words, they have a certain weight which is supposed to be a measure of the strength of the support that facts give to those things they are reasons for.
5. Reasons can be *pro tanto* or *prima facie*
6. Given their *pro tanto* or *prima facie* nature, reasons can either be aggregated in some way or conflict with one another, or they can be overridden or defeated by one another, etc.
7. Reasons are those things that can be given by a third party as a piece of advice about what to do or believe.
8. Reasons serve as inputs to correct deliberation and reasoning.

Now that I have delineated the concept of a normative reason, I will proceed by examining the question of what the relation *counting in favor of* could stand for or the truth conditions of those claims, and whether they could be incorporated into "the world of facts as revealed by science" (Harman, 2000, p. 79).

2.4 Ontological accounts of reasons

Now we can ask what makes claims that some fact is a reason to Φ[14] or that a fact counts in favor of Φ-ing true or false? There are two general positions regarding the answer to this question. One is to claim that the fact that something is a reason is a normative fact that exists independently of the mind or subject that responds to it. The other is to deny the latter and claim that the matter of which facts provide reasons is mind- or subject-dependent. The question under consideration can be put in terms of Euthyphro's dilemma. Is there a reason to believe, desire,

[14] Where Φ-ing could be the formation of some attitude or performance of an action.

concern, intend, value, judge valuable, etc. because there are some irreducibly normative facts, or are there facts about reasons because we believe, desire, have concerns, intend, value, etc.? (cf. Enoch, 2005, pp. 763-764). With this question in mind, I will proceed with Parfit's (2011a) characterization of the following two accounts of reasons.

2.4.1 Object-based theories of reasons

Parfit starts from the folk-psychological characterization of motivational states.[15] Like many others (Davidson, 2001; Smith, 1987; Williams, 1981), Parfit takes the word "desire" to refer to "any state of being motivated, or of wanting something to happen and being to some degree disposed to make it happen, if we can" (Parfit, 2011a, p. 44). The second important thing that Parfit emphasizes is that desires have contents, or in his terms "desires have *objects*, which are *what* we want" (ibid.). In this sense desires are like beliefs.

However, desires and beliefs are different types of mental states. The relevant difference between them is often cashed out in terms of the metaphor of *directions of fit* (Anscombe, 1957; Searle, 2001; Smith, 1987; however see Sobel & Copp, 2001). There are two relevant directions of fit in this context: *world to mind* and *mind to world*. Desires have a world to mind and beliefs a mind to world direction of fit. These notions explain the following idea. Desires and beliefs can have the same content. For instance, someone could desire rain to fall now, and someone else could believe that rain is falling now. However, despite the possibility of desires and beliefs having the same content their satisfaction conditions will be different depending on their direction of fit.

The desire for the rain to fall will be satisfied when the world conforms to the mind, that is, in virtue of the world being in accordance with the desire. The belief that it is raining now, on the other hand, will be satisfied if it is really raining, that is, when the content of the mind conforms to how the world really is now.

Objects of desires can include "acts, processes, and states of affairs" (Parfit, 2011a, p. 44). For example, we can desire to drink a beer, to be happy, to have a nice evening, to be with our loved ones, for our favorite sports team to win, and so on. Parfit also distinguishes between teleological or telic and instrumental desires. Desires are "*telic* when we want some event as an *end*, or for its own sake" (ibid.). On the other hand, "desires are *instrumental* when we want

[15] At this point Parfit (2011a) frames the debate solely in terms of practical reasons. As such, the following discussion will be framed in terms of practical reasons.

some event as a *means*, because this event would or might cause some other event that we want" (ibid.). Some examples of *telic* desires might be the desire for our children to be happy and safe, to live long and prosper, to avoid agonistic pain as much as possible, etc.

Instrumental desires serve as instruments for accomplishing some other more basic desires. For instance, we might want our children to graduate from good schools because that will allow them to live more satisfactory lives; we might desire to go to cinema because that will enable us to have a pleasant evening; we might want to learn a lot about science so that we can satisfy our need for understanding how the world really works, etc. Parfit mentions that we sometimes want some acts or events as ends, but also as means to some other end. In this respect, Parfit says that "[t]wo such events might be a thrilling search for some important truth, and, when we want to have a child, making love. When we decide to try to fulfill some telic desire, we thereby make this desire's fulfilment one of our aims" (ibid.). It is natural to suppose that instrumental desires form chains that are grounded on some telic desire. For example, someone could desire to work on Wall Street because she wants to become rich, and she might want to become rich in order to buy nice things. Furthermore, she might want to have nice things in order to be recognized in her community and she wants this, for example, because it would allow her to find a suitable spouse, and so on.

Parfit seems to think that all chains of instrumental desires will be grounded on some telic desire, but not necessarily the same one (ibid.). This also includes the converse, that some telic desire will be at the beginning of a particular chain of instrumental desires (see also Smith, 2004). Furthermore, from this point of view it seems to follow that there is some foundation according to which every chain of instrumental desires is based on some telic desire that is not based on any other chain of instrumental or telic desires. Whether this foundationalist view is plausible is not very important at this point of the discussion. However, what is important is that we have enough of background to introduce Parfit's distinction between object-based and subject-based theories of practical reasons.

The first horn of Euthyphro's dilemma is captured by theories that Parfit calls object-based theories of practical reasons. According to object-based theories of practical reasons, "there are certain facts that give us reasons both to have certain desires and aims, and to do whatever might achieve these aims" (Parfit, 2011a, p. 45). These theories are called *object*-based theories because, according to Parfit, "[t]hese reasons are given by facts about the *objects* of these desires or aims, or what we might want or try to achieve" (ibid.). Parfit adds that if we believe that "all practical reasons are of this kind, we are *Objectivist about Reasons*, who accept or assume some *objective* theory" (ibid.).

Furthermore, Parfit explains why object-based theories can be called *value-based* theories. The reason for this is the following:

> Object-given reasons are provided by the facts that make certain outcomes worth producing or preventing, or make certain things worth doing for their own sake. In most cases, these reason-giving facts also make these outcomes or acts good or bad for particular people, or impersonally good or bad. *(Parfit, 2011a, p. 45)*

According to Parfit, object-based theories claim that reasons for action are provided by objects or possible contents of our desires, that is, by facts that make some act or some outcome valuable for its own sake. Furthermore, we can add that negative reasons or reasons for avoiding something are provided by facts that make certain acts or outcomes in some way bad. Besides Parfit (2011a; 2011b), authors that could naturally be placed among proponents of object-based theories of practical reasons include, for example, Enoch (2011), Scanlon (1998), and Shafer-Landau (2003).

Object-given reasons are provided by facts about the event or act that could be an object of our desire. Parallel to the difference between telic and instrumental desires, Parfit introduces a distinction between telic and instrumental reasons:

> [R]easons are telic when they are provided by the facts that make some possible event good as an end, or worth achieving for its own sake. Such reasons are instrumental when they are provided by the fact that some event would have good effects, by being a means to some good end. *(Parfit, 2011a, p. 52)*

Parfit also distinguishes between intrinsic and extrinsic telic reasons. On the one hand "[t]elic reasons are *intrinsic* when they are provided by facts about some possible event's intrinsic properties or features, or what this event would *in itself* involve. We might have such reasons, for example, to want to make someone feel less lonely, or to see the sublime view from the summit of some mountain, or to understand how life or the Universe began" (Parfit, 2011a, p. 52). On the other hand, telic reasons are extrinsic when the reasons are provided by good-making facts about some "event's relation to other events" (ibid.). However, Parfit does not place too much stress on extrinsic as opposed to intrinsic telic reasons because "events would be extrinsically good by making some longer sequence of events, of which they were one part, intrinsically better" (ibid.).

To illustrate what has been said so far about object or value-based theories, we can give the following example. Let us suppose that harming other people by inflicting pain on them is bad. Then, according to the theories under consideration, the fact or facts that make harming

bad (such as causing insuperable pain to another person) is an intrinsic reason not to do it. In other words, the *intrinsic features* of pain provide intrinsic reasons to avoid pain or, in this case, to avoid hurting other people. Alternatively, to give a more positive example, let us suppose that discovering the truths of the universe has an intrinsic value. In that case, the facts that make discovering the truths of the universe intrinsically valuable, such as the feeling of happiness and satisfaction when a certain level of scientific understating is achieved, provide one with intrinsic reasons to want to or to try to discover the truths of the universe. Thus, in this kind of theory, the emphasis is on the *features* that *make* certain states of affairs *valuable*, and those features are reasons, or to be more precise, they provide reasons to want or to do things.

The basic idea of object-based theories, according to Parfit, seems to be that the value of certain facts is intrinsic to those facts in the sense that those facts are such that they make certain things valuable completely objectively, without reference to the subject who might find them valuable. Furthermore, the idea seems to be that they would still be valuable even if no one existed who could appreciate their value-conferring potential (see also Enoch, 2011; Shafer-Landau, 2003).

In terms of truth-conditions, object-based theories claim that statements about normative reasons refer to irreducibly normative facts and properties. This means that normative truths, such as that X has the property of being the right thing to do or the property of being what one ought to do, are irreducible and cannot in any way be connected to, for example, naturalistic facts about motivation (Parfit, 2006; 2011b, p. 486). According to this view, truths about reasons are *necessary*, and their status is often compared to mathematical and logical truths (Parfit, 2011a, p. 129; 2011b, pp. 307, 326, 489, 643, 746). Since it is normally thought that mathematical and logical truths are *discovered* through mathematical reasoning and reflection,[16] by analogy, the idea should be that normative truths about reasons are also true across all possible worlds and are discovered through reasoning and reflection on facts. Here is how Parfit phrases this point:

> Fundamental normative truths are not about how the actual world happens to be. In any possible world, pain would be in itself bad, and prima facie to be relieved rather than perpetuated. Similarly, even if the laws of nature had been very different, rational beings would have had reasons to do what would achieve their rational aims. As in the case of

[16] Or in Parfit's words: "We often *can* discover logical or mathematical truths merely by thinking about them" (Parfit, 2011b, p. 489).

> logical and mathematical truths, we can discover some normative truths merely by thinking about them. *(Parfit, 2011b, pp. 489-490)*

Parfit's development of an object-based theory of reasons is problematic from a naturalistic point of view. It seems to be a platitude about normative reasons that one of their main roles is to motivate, direct, or govern actions and beliefs (Korsgaard, 1986; Smith, 1994). In order for them to fulfill this important role, it seems that they need to be in an important way accessible and related to rational agents. Furthermore, if reasons have this motivational role, then it is natural to think of them as being dependent on the activity of a being who can respond to them, think about them, and act on them. By comparing claims about reasons with claims about mathematics, this important governing relation seems to be undermined. Normally, we do not conceive of the objectivity of mathematical statements as being dependent on the responses of agents. But then again, taken in their completely objectivistic guise, we do not take it that one of the essential features of mathematical truths is to govern action. This seems to be a big and an important disanalogy between the necessity of mathematical truths and the necessity of truths about normative reasons.

Nevertheless, Parfit (1997) does not seem to be moved by this objection. According to him, truths about what one should do or want are wholly independent from what one actually wants or is inclined to do. In addition, what one should do is what one should do, regardless of whether this fact actually motivates you or would motivate you should you be aware of the relevant normative fact. This hyper-objectivistic stance regarding reasons is, however, what creates a puzzle for this family of views. On the one hand, reasons are thought of as being provided by states of affairs, and that some state of affairs is a reason for something is supposed to be a completely mind-independent normative fact. On the other hand, such reasons should apply to and govern the actions and mental states of real-life agents. The puzzle is, first, how these mind-independent facts about what we should do have as outputs actions and attitudes that are paradigmatically mind-dependent, but nevertheless remain wholly mind-independent. Second, and more importantly, if reasons provide necessary truths, then the puzzle is how they come to be antecedently arranged, weighted, and fitted to apply to an arbitrary agent in a situation in which she needs to reach a decision. This puzzlement is nicely brought out by Christine Korsgaard in the following quote:

> Human beings, (…) need reasons. We cannot determine our beliefs or actions without them. And according to [object-based theories], when we look around us, we find them. But this seems like a mere piece of serendipity. The reasons are in no way generated by the problem that, as it happens, they solve; they just happen to be there when we need them. We need to

> make decisions, and lo and behold, we find around us the reasons we need in order to make those decisions, equipped with weights or strengths that will enable us to balance them up and arrive at a decision. *(Korsgaard, 2011, p. 6)*

If we grant that reasons are grounded in mind-independent facts, then it becomes mysterious how we get such a nice fit between the problems that we *happen* to need to solve and the pre-packed and pre-weighted reasons that *necessarily* solve them. Unless object-based theorists can provide some plausible explanation of how this magic fit came about between our reasons and who we as a matter of contingent fact are, we will be left, as Korsgaard writes, with a serendipitous view of normative reasons.[17]

In their most common guises, subject-based theories avoid this sort of puzzlement. Thus, in the next subsection I turn to the discussion of subject-based theories of normative reasons.

2.4.2 Subject-based theories of reasons

In contrast to object-based theories, subject-based theories are not oriented to the intrinsic features that make certain states valuable in themselves. Rather, they are more relational in character. Subject-based theories claim that:

> our reasons for acting are all provided by, or depend upon, certain facts about what would fulfill or achieve our present desires or aims. Some of these theories appeal to our actual present desires or aims. Others appeal to the desires or aims that we would now have, or to the choices that we would now make, if we had carefully considered all of the relevant facts. *(Parfit, 2011a, p. 45)*

It should be clear why the latter theories are called subjective and why, in terms of the Euthyphro dilemma, they represent the second horn. The claim is that reasons in some way depend on facts about agents and their desires (in the broad sense). Since Parfit construes subjectivist theories as being based in some way on an agent's desires, this group of theories can also be called desire-based theories. Among the influential authors that adopt some version of a subjectivist theory are Goldman (2010), Schroeder (2007), Smith (1994; 2004), and Williams (1981), but also authors coming from the Kantian and constructivist tradition, such as Korsgaard (1996) and Street (2008a).

From the above quote it can be discerned that the family of theories that fall under the title of subject- or desire-based theories will vary depending on how we interpret the phrase that

[17] Korsgaard's criticism might be further developed in different directions. In chapter 4 I will defend an evolutionary version of this criticism that was influentially developed by Sharon Street (2006).

'reasons depend on subjects'. For example, if we take the crude form and say that reasons are provided by facts that would fulfill our present telic desires, then we could get different predictions about what our reasons would be than if we interpret the phrase as saying that reasons depend on the desires that we *would* have after we engage in some sort of deliberation.

Thus, on the first interpretation the fact that I have a strong desire to eat a whole box of chocolates is a reason to eat them. However, it could be the case that were I to deliberate for a moment I would conclude that eating the chocolates now would be terrible for my health, such that I would lose the desire. In that case, the fact that I have a desire now would not be a reason to eat the chocolate. Since the existence of this sort of revision procedure seems plausible to me, in what follows I will construe desire-based theories as involving at least this sort of minimal check-and-revise procedure.

It is not easy to find a single coherent characterization of all subjectivist theories of practical reasons. Parfit goes through many views in his book, however, perhaps the most general characterization he provides is given in the following quotation:

> Subjectivism about Reasons: Some possible act is what we have most reason to do, and what we should or ought to do in the decisive-reason-implying senses, just when, and because, this act would best fulfil our present fully informed telic desires or aims, or is what, after ideal deliberation, we would choose to do. *(Parfit, 2011a, p. 64)*

As an exemplar of a subjectivist theory of practical reason, I will take Bernard Williams' theory of internal reasons as expounded in his seminal paper *Internal and External Reasons* (Williams, 1981; see also his 1995).

2.4.3 Internalism and the normativity of reasons

In his seminal article (1981, p. 101), Williams ask us to consider the following sentences: "There is a reason for A to Φ" and "A has a reason to Φ." We may wonder about the truth-conditions of these sentences. On object-based theories of practical reasons the truth-conditions of these sentences would include some properties of Φ-ing that make it intrinsically good and thereby count in favor of performing Φ. On Williams' internalistic account things are reversed, so that Φ-ing is favored or there is a reason to Φ because some desire from A's set of desires would be satisfied. Thus, Williams says that the sentence "A has a reason to Φ" is true iff A has some desire that will be served by his Φ-ing (1981, p. 101).[18] In his later work, Williams

[18] Williams (1981, p. 102) calls his interpretation *the sub-Humean model* because it is in the general spirit of

dropped the sufficiency condition and gave the following fuller explication of the statement that A has a reason to Φ:[19] A has a reason to Φ only if "A could reach the conclusion that he should Φ (or a conclusion to Φ) by a sound deliberative route from the motivations that he has in his actual motivational set – that is, the set of his desires, evaluations, attitudes, projects, and so on" (Williams, 1995, p. 35).[20] In contrast, externalist theories, in line with object- or value-based theories about reasons, would claim that whether A has a reason to Φ does not depend in any way on the agent's motivations.

Thus, on the subjectivist/internalist view, reasons are explained not by any intrinsic or irreducible features that acts or states of affairs might have, but by responses that those features might invoke in agents with certain profiles. And what profiles agents might exhibit depends on their motivational sets and what constitutes 'the sound deliberative route'.

Before I move on with the discussion of subjectivist theories, I want to address an important objection that Parfit raises against Williams' type of subjectivist theory of normative reasons. Parfit's objection can be stated as two interrelated points.

Parfit (2011b, sec. 84) complains that if we adopt Williams' reductive account of reasons then, in effect, we eliminate their normativity. Therefore, according to Parfit, that account of reasons cannot provide proper analysis of reasons. To illustrate this objection, Parfit offers the following line of reasoning:

> (A) Jumping into the canal is my only way to save my life.
> (B) Jumping is what, after rationally deliberating on the truth of (A), I am most strongly motivated to do.
>
> Therefore
>
> (C) As another way of reporting (B), I could say that I have most reason to jump. *(Parfit, 1997, p. 123)*

Parfit objects that (C), if it is a statement about normative reasons, cannot be just a restatement

---

Hume's view on practical reason, even though it is plausible that it does not capture Hume's actual view (for what might be Hume's actual view on practical reason see Millgram, 1995).

[19] This formulation is also present in his (1981) paper. Nevertheless, Williams continues to think that his formulation of the truth-conditions for reason-statements also provide a sufficient condition (Williams, 1995, p. 35).

[20] Briefly, Williams calls an agent's motivational set S and members of that set desires (Williams, 1981, pp. 102, 105), but, as should be clear from the quote, desires, as in Parfit's case, include all kind of pro-attitudes that an agent might have.

of (B), since (B)-type statements, according to Parfit, are not normative; they only provide an empirical or psychological prediction about what we would do or want after deliberation (cf. Parfit, 1997, p. 126). While reason-statements are supposed to tell us what we *should* do or rationally ought to desire.

However, this objection is not persuasive. As Parfit (ibid., 125) himself recognizes, Williams provides truth-conditions for statements about reasons in terms of *rational* deliberation or *sound* deliberative routes (cf. Roberts, 2005, p. 101). In this regard, (B) cannot be read as a purely non-normative statement. Whether I have a reason or most reason to jump does not depend on the bare casual force with which I form my desires. Rather, the normative status of those desires depends on the correctness conditions or standards of the processes that govern desire and belief formation. Since those standards are not simply casual, I may fail to satisfy them and therefore act irrationally. What Parfit and Williams might disagree about, here, is what constitutes the norms of rational (or sound) deliberation. However, this disagreement does not strip the notion of internal reasons of its minimal normativity.

Parfit might further object, and this leads us to the second point, that (B) cannot be what we mean by a purely normative statement such as 'I have a reason to Φ' since (B) is at least partly an empirical prediction about what we would be motivated to do. However, according to Parfit purely normative reasons cannot be defined in any other terms, especially non-normative terms. Here is how Parfit explains his view:

> It is hard to explain the concept of a reason, or what the phrase 'a reason' means. Facts give us reasons, we might say, when they count in favour of our having some attitude, or our acting in some way. But 'counts in favour of' means roughly 'gives a reason for'. Like some other fundamental concepts, such as those involved in our thoughts about time, consciousness, and possibility, the concept of a reason is indefinable in the sense that it cannot be helpfully explained merely by using words. We must explain such concepts in a different way, by getting people to think thoughts that use these concepts. *(Parfit, 2011a, p. 32).*[21]

According to Parfit, in order for (B) to be purely normative, the concept of rationality should be read as *substantive* rationality. However, substantive rationality cannot be expressed without

[21] In this respect Parfit echoes Scanlon's view on reasons: "I will take the idea of a reason as primitive. Any attempt to explain what it is to be a reason for something seems to me to lead back to the same idea: a consideration that counts in favor of it. 'Counts in favor how?' one might ask. 'By providing a reason for it' seems to be the only answer" (Scanlon, 1998, p. 17).

saying that "we must want, and do, what we know that we have most reason to want and do" (Parfit, 1997, p. 116), which "could be true even if, […] no amount of informed deliberation would in fact motivate [an agent]" (ibid., 101). Since Parfit uses the concept of a normative reason in this pure, non-psychological and irreducibly normative sense, he even thinks that he and Williams could not have normative disagreements, because Williams' claims about reasons and what ought to be done "are really psychological claims about how we might be motivated to act" (Parfit, 2011b, p. 452).

As we have seen, Williams' notion of an internal reason cannot be purely psychological or empirical since it essentially invokes norms of rational deliberation. Nevertheless, even if we grant that the concept of a normative reason is primitive it still does not follow that Williams' internalism is not about *normative reasons* (see Appendix on Williams in Scanlon, 1998). As Street points out, even if we grant that truths about reasons are irreducible and primitive, in the sense that we cannot define them in any other (especially not in non-normative) terms, it still does not follow that understanding the notion of a normative reason entails the falsity of internalism about reasons (Street, 2016). According to Street, one way in which we might acquire the concept of a normative reason

> is to point to a certain type of conscious experience with which we're all intimately familiar. The intrinsic character of this experience cannot accurately be captured or described except by invoking normative language—just as, for example, the intrinsic character of the experience of redness cannot accurately be described except by invoking color language—but that doesn't mean we can't locate for one another the type of experience in question by pointing to the kinds of circumstances in which those of us who are party to the discussion tend to have it. In other words, just as we point to the experience of redness by pointing to the kinds of circumstances in which we typically have it—for example, when looking at ripened strawberries or a fire truck—so we may point to what we might call the experience of 'to-be-done-ness' by pointing to the kinds of circumstances in which we typically have it—for example, when a car suddenly swerves toward us on the highway, or when we see a child in pain. *(Street, 2016, pp. 3-4)*

I do not want commit myself to the plausibility of Street's account of the way in which we acquire normative concepts. What is important for our present purposes is that this account shows us that having a concept of a purely normative reason does not necessarily imply anything about its underlying metaphysics. In particular, possession of the concept of a normative reason does not preclude the possibility of our reasons being fixed by sound or rational deliberative routes that start from our actual motivations. On the other hand, it does not

preclude the possibility of reasons being external, that is, fixed by completely mind-independent facts.

After we grant that internalist theories of the type provided by Williams can be interpreted as providing an account of the nature of normative reasons, the important question becomes: what constitutes the sound, or in other words, rational deliberative route? This question is important because what reasons one has will depend on how we construe the latter. Concerning this point, Williams writes that "[t]here is an essential indeterminacy in what can be counted a rational deliberative process" (Williams, 1981, p. 110).

Williams took it that the rational deliberative route includes rather thin norms of reasoning so that, in his view, it is largely a contingent fact what particular agent has a reason to do. In particular, Williams thought that the rational deliberative route would involve "at least correcting any errors of fact and reasoning involved in the agent's view of the matter" (Williams, 1995, p. 36). Hence Williams' famous gin and tonic example. Mary may have a desire to drink the stuff in her glass but she does not have a reason to do so because if her beliefs were corrected she would cease to desire to drink the stuff that is in the glass. Other examples, with the exception of causal means-ends reasoning, of how a person might come to the conclusion that she has a reason to act in some way are provided in the following quote:

> A clear example of practical reasoning is that leading to the conclusion that one has reason to Φ because Φ-ing would be the most convenient, economical, pleasant etc. way of satisfying some element in S, and this of course is controlled by other elements in S, if not necessarily in a very clear or determinate way. But there are much wider possibilities for deliberation, such as: thinking how the satisfaction of elements in S can be combined, e.g. by time-ordering; where there is some irresoluble conflict among the elements of S, considering which one attaches most weight to […]; of again finding constitutive solutions, such as deciding what would make for an entertaining evening, granted that one wants entertainment. *(Williams, 1981, p. 104)*

Whether some other norms or patterns of practical reasoning necessarily belong to an agent's motivational set is a matter of dispute (Korsgaard, 1986). For example, Williams did not think that moral considerations necessarily belong to an agent's motivational set. To illustrate this, he gave an example in which we suppose that he is a person who thinks that someone ought to be nicer to his wife:

> I say, 'You have a reason to be nicer to her'. He says, 'What reason?' I say, 'Because she is your wife.' He says—and he is a very hard case—'I don't care. Don't you understand? I really do not care.' I try various things on him, and try to involve him in this business; and I find

> that he really is a hard case: there is nothing in his motivational set that gives him a reason to be nicer to his wife as things are. *(Williams, 1995, p. 39)*

Williams adds that one can try to influence this kind of person using different means, such as by saying that "he is ungrateful, inconsiderate, hard, sexist, nasty, selfish, brutal, and many other disadvantageous things" (Williams, 1995, p. 39). But if nothing works, then according to Williams such a person would not have any reason to be nicer to his wife. Such persons, who would seem to have psychopathic traits, in the sense that they do not care about the feelings of other people and without a sense of regret take other people for granted (or use them in more devious ways), seem to be ubiquitous in our society (see e.g. Babiak & Hare, 2006; Hare, 1993). According to Williams, these kinds of people, if otherwise rational in their reasoning capacities, would not have any reason to follow moral prescriptions (such as to be nice to your spouse, not to hurt other people, to apologize when you do something wrong, etc.) that we normally take for granted.

Even though Williams (1995) is skeptical of this, some subjectivists would claim that prudential and moral norms necessarily belong to the motivational sets of every rational agent (see e.g. Korsgaard, 1996; Smith, 1994). If that were true, we could say that every rational agent necessarily has a reason to act morally because she could reach a reason to act morally from any motivational set she starts from. Or to be more precise, she would already have a reason to act morally because moral norms would be part of her rational deliberative route that governs and transforms her initial motivational set. The important thing to note here is that subjectivists are not *a priori* committed to claiming that only actual desires, whatever they might be, provide reasons to satisfy them or to act in some way.

Which norms constitute subjects' motivational sets and thereby constitute the norms of rationality is not important at this moment.[22] What is important is that subjectivists, according to my construal, endorse some kind of dispositionalist or even constructivist account of reasons (Smith, 1989; Street, 2008a). Thus, the general claim is that reasons are not provided by intrinsic properties of things that are encapsulated in the relation *counting in favor of*. The basic subjectivist idea is that the relation *counting in favor of* can be accounted for in terms of the relation between the structure of the rational agent and the environment in which she finds

[22] From the discussion in chapter 4 it will emerge that what reasons we have will largely depend on contingent facts that were fixed by evolutionary, developmental and cultural considerations. Thus, to a significant degree I agree with Williams that the norms of rationality that fix reasons cannot be determined on a priori grounds; rather they will reflect lots of contingent facts about us and our history.

herself. In a nutshell, the subjectivist's datum is that things are valuable or provide reasons because they speak to our desires and deepest concerns (Goldman, 2010), and in general because of what we value or would value under certain conditions.

2.4.4 The difference between object-based and subject-based theories

The difference between object- and subject-based theories can easily be misunderstood. The first difference that naturally comes to mind is that according to object-based theories reasons are objects or contents of mental states (desires, beliefs, etc.), while according to subject-based theories reasons are mental states themselves. However, this is not the right way to construe the difference. If that were the case the subjectivist theories would immediately look implausible, since they would not be able to account for the *counting in favor* of relation and how we normally conceive of it.

In fact, we normally talk about facts that are not about our desires as being reasons to do something or to believe something. Moreover, desires are normally not construed as relations that count in favor of something. At most the content of a desire or the fact that one has a desire that p is used as a grounding part of the *counting in favor of* relation.

Thus, rather than saying that on subjectivist theories desires are reasons, I will follow Alan Goldman (2010) and say that reasons are facts or states of affairs that can be objects of a person's desires. The crucial distinction between object-based and subject-based theories is ontological, in the sense that on both accounts reasons can be facts or states of affairs outside the agent, however they vary on what *makes* those facts reasons. On objectivist accounts they are irreducible normative facts, while on subjectivists accounts reasons lie "within valuing subjects" (Goldman, 2010, p. 28) or their dispositional properties as rational agents (Smith, 2004).

Besides the ontological difference in the latter sense, Parfit claims that subjectivist and objectivist theories can be differentiated by what those theories imply we have or do not have a reason to do or to want (Parfit, 2011a, pp. 46-47). According to Parfit, there are principled and deep disagreements between the implications of the two theories. Against the backdrop of this thought, he argues against subjectivist theories by claiming that they have some implausible consequences when it comes to the reasons we have and the reasons we *think* we have. Subjectivists disagree on this point and claim either that the intuitions about reasons that objectivists endorse can be accommodated by subjectivist theories or that we should revise our intuitions. In order to evaluate this concern, I will examine the type of argument that Parfit

offers against thinking that subjectivist accounts are even minimally plausible.

2.5 Subjectivism and its implications

2.5.1 The agony argument

Parfit starts his objection with the so-called *agony argument.* The argument takes for granted that we have decisive or at least sufficient reasons to want avoid all future agony. That is the datum of the argument. The argument is the following:

> Suppose that, in Case One, I know that some future event would cause me to have some period of agony. Even after ideal deliberation, I have no desire to avoid this agony. Nor do I have any other desire or aim whose fulfilment would be prevented either by this agony, or by my having no desire to avoid this agony. Since I have no such desire or aim, all subjective theories imply that I have no reason to want to avoid this agony, and no reason to try to avoid it, if I can. *(Parfit, 2011a, pp. 73-74)*

The idea is that according to subjectivist theories it is always possible not to have a desire to avoid future agony even if one were completely rational and rationally deliberated about the issue.[23] Therefore, according to Parfit, subjectivist theories are false (Parfit, 2011a, p. 76).

Against the argument, one could argue that even if it is true that it is logically *possible* that after ideal deliberation some agents would still lack the desire to avoid all future agony, this would not be true of the *actual* rational agents. However, Parfit has a retort to this line of reasoning:

> [E]ven if there were no such actual cases, normative theories ought to have acceptable implications in merely imagined cases, when it is clear enough what such cases would involve. Subjectivists make claims about which facts give us reasons. These claims cannot be true in the actual world unless they would also have been true in possible worlds in which there were people who were like us, except that these people did not want to avoid all future agony, or their desires differed from ours in certain other ways. So we can fairly test subjective theories by considering such cases. *(Parfit, 2011a, pp. 76-77)*

In this quote we can see that Parfit presupposes that certain claims about reasons need to be true necessarily (in all possible worlds similar to ours) in order to be true in the actual world. Wanting to avoid all future agony seems to be Parfit's paradigmatic example.

[23] Parfit offers other similar examples such as the future Tuesday indifference example (see chapter 4 in Parfit, 2011a). According to this example, we care what happens to us on every day except for Tuesday that is to come. The reasoning of this thought experiment is the same as above, in the agony argument.

However, that just seems to be Parfit's bias, driven by his views or intuitions about what reasons there are and what kind of theory accounts for them. It seems to me that a 'subjectivist' could give at least two possible responses. A subjectivist may consistently claim that claims about reasons are contingent since they depend on our rational dispositions to treat certain facts as reasons. But she does not need to claim that there will always necessarily be a fact or state of affairs that will count as a reason for something in all possible worlds for all rational agents. We can compare this idea with Williams' claim that there will not always be a definite answer to the question of what a person has a reason to do:

> [I]t is unclear what the limits are to what an agent might arrive at by rational deliberation from his existing S. It is unclear, and I regard it as a basically desirable feature of a theory of practical reasoning that it should preserve and account for that unclarity. […] Practical reasoning is a heuristic process, and an imaginative one, and there are no fixed boundaries on the continuum from rational thought to inspiration and conversion. […] There is indeed a vagueness about 'A has a reason to Φ', in the internal sense, insofar as the deliberative processes which could lead from A's present S to his being motivated to Φ may be more or less ambitiously conceived. But this is no embarrassment to those who take as basic the internal[24] conception of reasons for action. It merely shows that there is a wider range of states, and a less determinate one, that one might have supposed, which can be counted as A's having a reason to Φ. *(Williams, 1981, p. 110)*

As should be clear from the quote, Williams thinks that what reasons we have cannot always be determined on a priori grounds, and that the account that captures and accounts for this aspect of normative reasons is actually better than alternative accounts.

At this point Parfit could retort that having both a reason and a recognition of this reason to want to avoid all future agony is so basic that it cannot depend on contingent opportunities and possibilities for practical reasoning. It might seem like Parfit's point holds, that is, that having a reason to want to avoid future agony cannot be desire-based if we allow that it is logically possible that after ideally rational deliberation an agent can fail to have a desire to avoid future agony. At this point, some subjectivists dig in their heels and defend the logical possibility. For instance, Street (2009) argues that if a person really after ideal rational deliberation still does not want to avoid all future agony, then such a person really would not have a reason to want to avoid all future agony. Furthermore, Street (ibid.) argues that this consequence, in fact, goes in favor of subject-based theories because it makes sense of the logical possibilities that thought experiments (future agony, future Tuesday indifference, etc.)

[24] In our present terminology we could say subject-based conception of reasons.

pertain to demonstrate. In fact, we might ask ourselves, what reason could a completely rational person have to want to avoid all future agony, if after rationally considering all the relevant facts and possibilities, she still does not think that she has a reason to want avoid all future agony or just lacks that desire? It is not clear what answer could we give to this question if we persist in believing that it is logically possible that after ideal deliberation we could still lack the desire to avoid all future agony (however, see Parfit, forthcoming).

Another way in which a subject-based theorist might respond is to accept the intuition that it is necessary that we have a reason to want to avoid all future agony, but to reject the possibility that after ideally rational deliberation one could fail to have a desire to avoid all future agony. To see how this could be done, we need to remember Parfit's claim that reasons "cannot be true in the actual world unless they would also have been true in possible worlds in which there were people who were like us" (Parfit, 2011a, p. 77). If we take it for granted that we look into possible worlds where there are only 'people like us', then it becomes plausible to argue that given who we are and our nature as rational beings it is not possible for us to be rational and fail to have even the slightest motivation or desire to avoid all future agony (Smith, 2009).

Thus, one could argue that given the fact that on subject-based accounts reasons supervene on the principles of rational deliberation and our *actual* nature, it is not possible that after ideally rational deliberation one would not have *any* desire to avoid all future agony. Furthermore, it is open for a more liberally inclined subject-based theorist to argue that even though it is *logically possible* that there is some rational being who after ideal deliberation would fail to have the relevant desire, that being would be totally unlike ourselves, and would not present a problem for subject-based theories because, given our actual natures as rational human beings, it is not possible for us to be ourselves and to lack even the slightest desire to avoid all future agony.

However, it could be argued that people as a matter of fact fail to always desire to avoid all future agony. In fact, Parfit seems to think that it is not true of actual people that they always have such a desire:

> Many people care very little about pain in the further future. Of those who have believed that sinners would be punished with agony in Hell, many tried to stop sinning only when they became ill, and Hell seemed near. And when some people are very depressed, they cease to care about their future well-being. *(Parfit, 2011a, p. 76)*

However, from a subjectivist point of view, there are two plausible ways to defend the subjectivist position. One is to question the rationality of the people in the example. Depressed people, at least, are often taken as paradigmatic examples of persons whose rationality is to

some point diminished (Smith, 1994). Thus, it is not clear how good a case Parfit presents in the above quote.

Second, it might be questioned whether it is really the case that these people really do not have *any*, even the *slightest* desire to avoid pain in the far future. It is important to emphasize that for a subjectivist about reasons to maintain her position, it is sufficient to claim that after ideal deliberation an agent would have *some* desire to avoid all future agony, but not necessarily an overriding desire to do so (cf. Sušnik, 2015). Parfit's examples and intuitions about logical possibilities do not demonstrate or show conclusively that actual people lack the *pro tanto* desire, or *a fortiori* that they would lack such a desire after they rationally deliberated about the issue.

In what follows, I will address another argument offered by Parfit that seems to me to be much more serious, because it claims that subjectivist theories are not coherent and thereby should be discarded. This is the so-called *incoherence argument.* In the next section, I will present the argument and try to show why it does not undermine subject-based theories of practical reasons.

2.5.2 The incoherence argument

The incoherence argument[25] consists of a statement that plausibly describes a whole range of subjectivist theories of reasons and a second statement that expresses the conditions that need to be satisfied in order for the first statement to be true. In particular, the punch line of the argument, according to Parfit, is that a subjectivist cannot accept the validity of the second statement and that is what makes his or her position incoherent. So, here is the first (M) statement of Parfit's argument:

> (M) what we have most reason to do is whatever would best fulfil, not our actual present telic desires or aims, but the desires or aims that we would now have, or would want ourselves to have, if we knew and had rationally considered all of the relevant facts. *(Parfit, 2011a, p. 93)*

Parfit adds one more, seemingly harmless, condition that seems reasonable from an epistemological perspective. That is the statement (N):

[25] Parfit's incoherency argument needs to be distinguished from Michael Smith's *incoherence argument,* as labeled by Shafer-Landau (1999, as cited in Smith 2004, essay 2).

> (N) when we are making important decisions, we ought if we can to try to learn more about the different possible outcomes of our acts, so that we can come to have better informed telic desires or aims, and can then try to fulfil these desires or aims. *(Parfit, 2011a, p. 93)*

Parfit then goes on to claim that (M) and (N) could only be true if statement (O) is true as well:

> (O) these possible outcomes may have intrinsic features that would give us object-given reasons to want either to produce or to prevent these outcomes, if we can. *(Parfit, 2011a, p. 93)*

Parfit (2011a, p. 93) gives the example of juries, which plausibly ought to consider the relevant facts, which would in effect give them reasons to believe that the accused person is guilty or innocent and thereby enable them to reach a final verdict. Similarly, Parfit claims, "when we are deciding which outcomes we shall try to bring about, we ought in important cases to try to discover, and rationally consider, what these outcomes would be like" (ibid.).

Everything that has been said so far seems reasonable from the perspective of subjectivist theories. Nevertheless, Parfit maintains that a subjectivist who accepts (M) and (N) cannot consistently accept (O), because (O) is precisely what object-based theories accept and what subject-based theories (should) reject. Since (N) presupposes (O), Parfit claims that subject-based theories cannot accept (N) (Parfit, 2011a, p. 94). Moreover, according to Parfit, subjectivists cannot accept either (M) because

> [i]f (O) were false, as Subjectivists claim, we would have no reason to believe that what we have most reason to do is whatever would best fulfil, not our actual present desires or aims, but the desires or aims that we would now have if we had rationally considered all of the facts about the possible outcomes of our acts. And if these facts could not give us reasons to have these desires or aims, we would have no reason to accept (M). We would have no reason to believe that these better informed desires or aims have any higher reason-giving status, or are desires or aims that we have more reason to try to fulfil. *(Parfit, 2011a, p. 94)*

Parfit's argument seems to be that subjectivists are committed to (M) and (N). They in turn presuppose (O). However, subjectivists do not accept (O). Therefore, their position is incoherent.

To evaluate Parfit's argument it is important to notice that according to subjectivists (M) is an ontological claim about the nature of reasons. The claim is supposed to account for the *counting in favor of* relation, and is not strictly related to the specific grounds of that relation. So to use Williams' model again, we can say that the fact that p counts in favor of Φ-ing iff there is a sound deliberative route that could lead one from the fact that p to Φ-ing. We are

explicating the concept of *counting in favor of* in terms of the concept of sound or rational deliberative route.

Statement (N) has a more epistemological or methodological flavor, since it is about how one should behave and think when one tries to reach an important decision. The role of statement (O) is to explain why we would use something like (N) in order to reach our reasons for action. Moreover, Parfit seems to presuppose in (O) that the explanation for why we use (N) in reaching decisions must rely on the existence of object-given reasons that are provided by intrinsic features of certain acts or events.

However, a subjectivist does not have to deny that intrinsic features of events and acts can be the grounds of reasons. The only thing she needs to deny is that what *makes* those facts reasons is their intrinsic nature. In other words, a subjectivist can claim that what makes those features count in favor of something is that they would lead a rational person from considering those features or facts to a decision to do something. Thus, it seems to me that (O) does not have to be true in order for (N) to be true. It is enough that (O') holds:

> (O') possible outcomes may have intrinsic features that would give us *subject*-given reasons to want either to produce or to prevent some outcome.

If some features of possible outcomes would give us reasons (which in this context means subject-based reasons) to want or to produce those outcomes, then we would have an explanation for why it could often be wise to follow a methodological principle such as (N).

The question now is whether (O') could explain (M). I think that (O') can explain (M), but as a consequence it will make (M) an analytical statement. If (O') holds then what gives us a reason to believe that what we would want after we have subjected our motivational set to ideal process of deliberation is what we have most reason to do is the fact that the two are the same, that is, that there is the conceptual link between *having a reason* and *desiring after ideal deliberation.*

Whether this is a problem for the subjectivist still needs to be investigated. Parfit (2011b) has arguments against what he calls analytical subjectivism but they do not purport to show that analytical subjectivism is incoherent (see also Parfit, 2011a, pp. 72-73). However, at least *prima facie* it does not seem to be incoherent to claim that what gives us a reason to believe that (M) is true is the fact that (M) explicates the concept of a reason. This point holds, it seems to me, if we bear in mind that (M) or something like it, such as Williams' concept of a sound deliberative route, intends to explicate the concept of a reason understood by the phrase *counting in favor of.* The difference between object-based and subject-based theories of reasons,

as I see it, is the way in which they account for the *counting in favor of* relation and not the more substantive question of what facts (states of affairs, or their features, etc.) exactly count in favor of what. The answer to the latter question will depend on how we construe the notion of a sound deliberative route (if we are subjectivists) or on more direct intuitions about the real value of things (if we are objectivists) (see Smith, 2009).

Parfit may concur with the above line of reasoning, since, when giving the incoherency argument, he seems to presuppose only what he calls subjectivist theories that are *substantive* with regards to what reasons we have, and not merely analytical. To make substantive claims about reasons, according to Parfit, one "must use the words 'reason', 'should', and 'ought' in the indefinable, normative senses" (Parfit, 2011a, pp. 72-73; see also section 5.2.1. above).

Now the question is, if subjectivists accept that the notion of a reason is primitive, in the sense that it cannot be defined in terms of, for example, a rational deliberative route, does this make their theories incoherent? Can analytical subjectivists alone avoid the incoherency argument? I do not think this is the case. Let me explain why.[26]

Even if we think that the concept of a reason is normatively irreducible, in the sense that it cannot be defined in any other terms, it still does not follow that statement (M), or some version of it, does not provide truth-conditions for the claim that there is a reason to do something. To simplify the point, we can maintain that the statement "There is a reason to Φ" is extensionally equivalent to the statement "there is a rational deliberative route that could lead one to Φ." Claiming that these two statements have extensional truth-conditions is not claiming that the concept of a reason reduces to or has the same meaning as the concept of a rational deliberative route.[27] Thus, a person who is competent regarding the concept of a reason does not have to *a priori* recognize, simply on the basis of his competency with the concept of a reason, that all that is captured with the concept of a reason is also captured by the concept of a rational deliberative route.

One explanation for this possibility is the fact that the concept of a rational deliberative route is not committed to any special view on what reasons there are (Smith, 2009). The only thing that a rational deliberative route, construed in subjectivist terms, needs to presuppose is

[26] The line of defense that will follow is similar to, and obviously influenced by (Street, 2016). See also section 2.4.3 above.

[27] For example, Christopher Peacocke gives the following criterion for when two concepts are distinct: "Concepts C and D are distinct if and only if there are two complete propositional contents that differ at most in that one contains C substituted in one or more places for D, and one of which is potentially informative while the other is not" (Peacocke, 1992, p. 2).

that, whatever reasons there are, they are reasons because of some connection with rational agents, and not because of the intrinsic value of certain state of affairs. The bottom line of the propounded view is that objectivists and subjectivists can share a concept of a reason, they can even claim that this concept is normatively irreducible, but they can still disagree on substantive issues regarding why certain reasons are reasons for certain agents or what makes them reasons, etc. (see Street, 2016).

Now we should be able to see why Parfit's incoherency argument also fails even if we interpret (M) as non-analytical. It fails because Parfit presupposes that when we use the concept of a reason as a primitive concept then we are committed to an object-based grounding of the relation *counting in favor of*. However, this need not be the case since we can share our notion of a normative reason (at least pre-theoretically) without sharing deep ontological commitments about the extensions of our concepts.

This point can be further illustrated with an example (see e.g. Hardin, 1988). Let us suppose that there are two persons, Joe and Mary. Mary was raised by parents who endorsed the philosophy according to which colors are objective and, we might say, intrinsic features of objects. In other words, they were realists about colors. Joe, however, was raised by parents who endorsed a response-dependentist ontology. According to them, colors are not intrinsic features of objects, instead they are dispositions of objects that could in certain circumstances induce a color experience in subjects that are perceptively exposed to those objects. Regardless of their background theories, when Joe and Mary meet and talk about colors they perfectly well understand each other; from their perspectives, they are both competent in using color concepts. Thus, when Joe asks Mary "Could you give me that red cup?," Mary responds by giving him the red cup. When Mary says "The color of this house is hideous," Joe agrees because he does not like houses painted in glaring green and red either.

It seems safe to say that Mary and Joe are competent in applying color concepts, and most of the time when they talk about colors they are talking about the same things. The only situation in which they would not agree is the situation in which the nature of color is discussed. Because of their different upbringings, Mary thinks that colors are intrinsic features of objects and Joe thinks that colors are response-dependent properties of objects. Thus, whether we think that background ontological theories should partly constitute a concept and our competence with it or not, we can safely say that at least pre-theoretically, that is, before the issue of the ontology of colors comes up, Mary and Joe share the same concept of a color.

To return to our discussion about the concept of a reason, we can say that the irreducibility, indefinability, or primitiveness of the concept of a reason does not commit us to any special

ontology of reasons, in the same way that the indefinability and primitiveness of particular color concepts does not commit us to any special color ontology (cf. Street, 2016). If we grant that the platitudes about normative reasons given in table 1 are something that theories of reasons should be able to accommodate, then, at least *prima facie*, it seems that object-based and subject-based theories are on the same pre-theoretical footing.

This line of reasoning should enable us to see that even if we construe Parfit's claim (M) as substantive, we can still say that what gives us reason to believe that our reasons are fixed by (M) is (O'), because the main issue between objectivists and subjectivists about reasons is the ontology of reasons and not which reasons exist.

2.5.3 Why idealize?

At this point I want to explore one more line of thought that might be propelling Parfit's intuitions behind the incoherency argument. This examination will lay the ground for what will follow in the next chapter.

One rationale for Parfit's contention that only statement (O) can explain why statements (M) and (N) hold is the view that if normative reasons are not mind-independent and provided by intrinsic features of things, then there would be no point in introducing idealization conditions into the picture about reasons.

David Enoch (2005) forcefully argues for this contention. The worry is that if we cannot give some kind of non-*ad hoc* reason for introducing idealizations into subjectivist accounts then we should suppose only that what explains idealization or reasons for introducing idealization into our account is the existence of object-based reasons. So unless we can show in a non-arbitrary way why idealization should be an integral part of our account of reasons, the threat is that subjectivist theories (that endorse some version of idealization) would be dangerously unstable. To see why, let us consider the following example.

Normally we introduce idealization conditions when they are needed for some kind of reliable tracking of facts. Enoch gives an example:

> you want to know who is taller, myself or my wife. Having a look seems like a good idea, but of course not just any look will do. What you want is to have a look from a proper angle, from up close, when my wife is not wearing heels. *(Enoch, 2005, p. 762)*

The reason why, in this kind of case, idealization is a good method of proceeding is because otherwise we might be misled. In particular, "if you are much closer to me than to my wife, the suggested epistemic procedure will fail; it will not be a reliable indicator of the relevant fact"

(Enoch, 2005, p. 762). When we are in non-optimal epistemic position when making a judgment of some kind, then idealization helps us to correct those epistemically poor circumstances. Here it is important to note that this kind of epistemic idealization makes sense because we are trying to track some fact that obtains independently of the tracking procedure.

In the practical domain, Parfit's statement (N) has a natural explanation if we suppose that (O) is true, namely if we suppose that there are facts that are worth discovering for their own sake. To rephrase the point just made, (N) as a methodological procedure makes sense if what it tracks is a procedure-independent fact.

However, as already noted when discussing Parfit's argument, this answer is not available to a subjectivist that construes the truth-conditions of reason claims as involving some kind of idealization condition. The reason for this seems to be clear: according to the statement (M), reasons are provided by what a rational deliberator would desire, decide, or aim to do. Thus it is not clear why we would want to idealize in order to get our reasons in order when what we are trying to get a grip on is not independent of our procedure for getting it. That is, it seems that whatever we have reasons to do will depend on what we desire to do after some kind of rational deliberation; but if that is so, then there is no point in trying to idealize because whatever output we get after the process of idealization is complete would be *the correct* output. This would seem to follow from the fact that there is no idealization-free fact that we are trying to get a grip on (cf. Enoch, 2005, pp. 764-765).

Thus, the pertinent question remains: if someone is a subjectivist about reasons why should she idealize? After evaluating three possible answers as to why a person like our subjectivist would want to idealize and showing their inadequacy, Enoch contends that someone would be justified in using idealization accounts, on a subjectivist construal, if it could be shown that the standard practice of idealization as a procedure of tracking independent facts and possibly the theory behind it is "badly mistaken, confused, or even incoherent" (Enoch, 2005, p. 785). If this could be shown, then the subjectivist could "suggest their idealizing view as the best reconstruction of the relevant discourse" (ibid.). In our case this would include showing that there is something badly mistaken with Parfit's object-based theory of practical reasons since this group of theories naturally explains why one would want to use idealization procedures in accounting for the reasons we have to act upon.

The last point can be made clearer through an analogy with dispositionalist theories of color. Intuitively we think, and moreover perceive color as an intrinsic feature of objects. For example, the desk in front of me is brown. The brownness of the desk seems to be an intrinsic property of the table. However, at least since Galileo this intuitive idea has been challenged.

John Locke (1690/1995) developed the idea that secondary qualities such as colors are features of objects that have a *disposition* to invoke, in subjects like us, a certain color-experience. Now, contemporary psychophysics tells us that colors might not be intrinsic features of objects (Giere, 2006). Given the empirical data, even though we intuitively think that colors are intrinsic features of objects, we can now justifiably claim that they are not. It is now completely plausible to claim that the brownness of this table in front of me, for example, is not an intrinsic feature of the table, but, for example, a dispositional property to invoke, in a perceiving subject, an experience of the table as being brown.

A similar thing can be said about normative reasons. Nevertheless, to be justified in saying this, it seems that we need to first show that there is something wrong with object-based theories or that an independently plausible argument can be given to the effect that subject-based theories can account for why we would want to use idealization in trying to determine what reasons for action we have. In the next two chapters the focus will be on discussing and tackling these issues.

2.6 Summary

In this long chapter my aim was to give an introduction to the topic of reasons. In this respect, I distinguished between motivational and normative reasons. Since the main aim of this thesis is to discuss the nature of normative reasons, in the rest of the chapter I limited the discussion to normative reasons. Following Parfit (2011a), I distinguished between two accounts of normative reasons: object-based and subject-based theories. I argued that from a naturalistic perspective, endorsing subject-based accounts makes more sense. Thus, in the rest of the chapter, I discussed prominent objections to subject-based theories and examined how they could be answered. I closed the chapter by discussing the 'why idealize' objection to a plausible type of subject-based theory given by Enoch (2005), in order to introduce the theme of the next two chapters.

3 Idealization and response-dependence

3.1 Introduction

In the previous chapter I wrote that the motivation behind Parfit's (2011a) *incoherence argument* might reflect Enoch's (2005) objection that subjectivists who essentially invoke some kind of idealization condition might have a problem justifying the need for idealization if they do not believe that normative facts are subject- or mind-independent. In this chapter I will argue that the response-dependent conception of normative reasons can provide a justification for idealization. I will argue by analogy with the case of color perception. To draw the analogy, it is enough to show that even if there are good grounds for thinking that color is some kind of response-dependent property, in the sense of being subject-based, this would not be incompatible with the introduction of idealization into the truth conditions of color ascription. I contend that, for the latter, it is enough to show that response-dependentist accounts can make a distinction between appearances and reality. My argument is that the same answer could be supplied against the incoherence argument, as interpreted in the light of Enoch's (2005) article.

I will proceed as follows. First I will introduce some considerations that indicate why it is a good idea for a subject-based account of reasons to accommodate the idea of idealization. In the process I will review some of Enoch's (2005) proposals for how a subjectivist might justify idealization. Then I will introduce the case of color perception and argue that similar considerations can be adduced in favor of a response-dependentist conception of normative reasons. I will close with some remarks concerning the relation between commonsensical account of reasons and the need for revision that response-dependentism potentially introduces, which will serve as an introduction to the topic of the next chapter.

3.2 Reasons for idealization

If someone is a subjectivist or internalist about reasons then the natural question that arises is why one should be fastidious about which desires, preferences, and concerns provide reasons and which do not. The natural answer is that we want to preserve our intuitions about, for example, which desires provide reasons and which do not. We tend to intuitively think that just because Mary has a desire to drink from the glass in front of her she does not really have a normative reason to do so (if the glass contains petroleum). We tend to think that were Mary

aware of the fact that the glass contains petroleum she would lose the desire to drink from it, and would see that her original desire did not give her a reason to drink from the glass.

Philosophers have devised a battery of examples to show that not every desire that one might actually have provides a reason for trying to satisfy it. The examples usually include what can be called defective desires (for a taxonomy of defective desires see Heatwood, 2005). These include having a desire to turn on every radio one notices (Quinn, 1993), having a desire to eat a saucer of mud when one sees it (Anscombe, 1957), having a desire to avoid all future agony except if it is Tuesday (so-called future Tuesday indifference (Parfit, 1984; 2011a)), having a desire to spend my life counting blades of grass (Rawls, 1971), etc.

Not everyone agrees on which of these examples involve desires that after a suitable idealization would stop counting as providing a reason for action. For example, Goldman (2010, section III) seems to think that having a desire to dedicate your life to counting blades of grass does not provide a real reason to actually go through with that plan. On the other hand Rawls (1971) thinks that having that kind of desire can provide a person with a reason to structure her life in accordance with it.[28] Whatever the final verdict on which desires, preferences, concerns, etc. provide reasons for action, a subjectivist wants to accommodate the intuition that not every whim, urge, or desire that surfaces in a person's mind provides a reason for performing some action or adopting some attitude.

If we disregard the details about which substantive desires provide reasons for action, the usual way for a subjectivist to develop her account is to introduce some counterfactual conditions by which we can check whether the desires under consideration survive some revision procedure. The procedure usually includes reference to the impact of true information and other coherence criteria on a person's set of desires (see Brandt (1979, chapter 6); Goldman (2010, chapter 2); Lewis (1989); Railton (1986); Smith (1994; 2004); Sobel (2009); Williams (1981; 1995)).

However, Enoch (2005) poses a general problem for a subjectivist who wants to introduce a procedure for desire evaluation that relies on some kind of idealization condition.[29] Enoch's claim is that if you want to be a subjectivist about reasons then the default position is to accept

[28] Indeed, the way Rawls describes the case it seems like the person from his example is completely reasonable and therefore his conception of a good life based on his desire also seems to be reasonable. According to Rawls' description of the example, the person "is otherwise intelligent and actually possess unusual skills, since he manages to survive by solving difficult mathematical problems for a fee" (Rawls, 1971, p. 432).

[29] Sobel in his (2009, footnote 3) lists other authors who raise similar objections to subjectivist theories of normative reasons.

that every actual desire provides a reason for action (Enoch, 2005, p. 760; see also Sobel, 2009). However, subjectivist theories that adopt the latter view have intuitively implausible consequences. Their implausibility stems from the implication that, for example, if you have a desire to eat a saucer of mud when you see one or to prefer a greater amount of pain later as opposed to a small amount of pain now then those preferences or desires would provide you with a reason for satisfying them. This is something that many authors want to avoid as implications of their accounts (Goldman, 2010; Parfit, 2011a; Smith, 1994; Sobel, 2009; Quinn, 1993).[30]

Thus, the question is: what is the rationale for introducing idealizing conditions? According to Enoch (2005, pp. 761-762), 'the natural answer' "would be to claim that the relevantly ideal conditions are the conditions needed for a reliable tracking of the relevant facts." By way of illustration, Enoch provides two examples. The first example involves the reliable tracking of time:

> Suppose that you want to know the time. Looking at a watch seems like a good idea. But, of course, looking at your watch may not be such a good idea. This depends on whether your watch keeps reasonably accurate time. What you want, then, is to have a look at a good watch. An ideal watch would be great, but we can settle for one that is less than ideal, so long as it is close enough. So we require, say, that the batteries in your watch be at least almost fully charged. *(Enoch, 2005, p. 762)*

The other example involves measuring sizes:

> Or consider this: you want to know who is taller, myself or my wife. Having a look seems like a good idea, but of course not just any look will do. What you want is to have a look from a proper angle, from up close, when my wife is not wearing heels. *(Enoch, 2005, p. 762)*

In these examples, according to Enoch, it makes sense to use idealization because our current epistemic situation might not reliably indicate what really is the case.

> If the watch is not reasonably accurate, or if you are much closer to me than to my wife, the suggested epistemic procedure will fail; it will not be a reliable indicator of the relevant fact (the time, or my and my wife's relevant height). And this, of course, is one good rationale for idealization: idealization (or its approximation) is called for whenever an actual procedure is fallible in ways (partly) corrected for by the idealization. *(Enoch, 2005, p. 762)*

[30] See Schroeder (2007) for an endorsement and defense of an actualist desire-based theory of reasons.

Rather than discussing Enoch's examples in detail I will introduce a third, more intuitive example to illustrate what seems to me to be Enoch's (2005) paradigmatic explanation of the 'natural answer' regarding the introduction of idealization. It seems to me that for Enoch (2005), the Müller-Lyer illusion would provide a paradigm case in which idealization is called for. In the Müller-Lyer illusion a naïve perceiver might think that she has a reason to believe that the two lines are of unequal length. But we can point out to her that if she *were* to measure the two lines she would see that they are actually not of the same length. In this case, we have a clear rationale for why some counterfactual condition defeats an apparent reason to believing something to be the case. Namely, by introducing better epistemic conditions one sees what really is the case. So the introduction of counterfactual conditions (in which we are in a better epistemic situation) makes sense because those conditions enable us to track what is the case *independently* of how we perceive things.

Enoch's (ibid.) argument is that this natural rationale for introducing idealization is not available to a subjectivist because according to the subjectivist there are no agent-independent facts that our reason-claims track. For a subjectivist a fact counts in favor of something because it indicates how her informed desires (cf. Lewis, 1989) or rational desires that reflect her deeper concerns (cf. Goldman, 2010) might be satisfied, and not because they track a pre-established normative order of things.

The second reason, according to Enoch (2005), why a subjectivist would want to count as reasons those desires that are accepted in ideal conditions is because she wants her account to be extensionally adequate when it comes to pre-theoretical intuitions about what reasons there are. However, just wanting to preserve extensional adequacy seems to be *ad hoc* as a rationale for introducing idealization in a subjectivist account because it does not offer any independent reason for accepting the subjectivist theory. Objectivist theories are also, supposedly, extensionally adequate and furthermore on Enoch's (ibid.) view they offer a 'natural' answer to why we should idealize; it is because idealization enables us to track *independent* facts. So on that score their being extensionally adequate is not *ad hoc*.

The third possible answer that Enoch (ibid., pp. 769-778) examines is the possibility that our justificatory practices concerning reasons and value actually reflect the need to idealize in circumstances when we think about our normative reasons. This certainly seems to be the case. Williams' (1981) gin-and-tonic example is a paradigmatic case that reflects our practices about reason claims. However, this will not help our subjectivist because the objectivist can accommodate this example and still provide Enoch's 'natural' answer as to why one should idealize in determining our reasons. Moreover, Enoch (2005, p. 774) claims that what explains

our need to idealize in ordinary justifications of our reason-claims is the fact that idealization is "conducive to the reliable tracking of an independent order of value-facts." Furthermore, why idealization explains it in this way is exactly because our commonsensical justificatory practice is committed, albeit implicitly, to the existence of agent-independent normative facts (ibid., p. 786; see also Sobel, 2009, pp. 341-342).

To illustrate the last point that Enoch (2005) makes we can use two intuitive examples that were used as counterexamples to Williams' internalist/subjectivist theory of reasons. Williams (1981; 1995) argues that what one has a reason to do depends on his or her preexisting motivations that are contingently given and to that extent, one can add, arbitrary. As a consequence of such an account the following case seems to be possible: a man abuses and molests his wife without having a reason to stop. We could think that that person *should* stop and that it would be *better* if he stops abusing his wife. Williams (1995, p. 39) says that there are many things we could say to that person, for example, "that he is ungrateful, inconsiderate, hard, sexist, nasty, selfish, brutal," but we would not be in a position to say that he has a reason to stop. According to Thomas Scanlon, the fact that we can say all these things to the abuser indicates that we are "accusing him of a kind of deficiency" (Scanlon, 1998, p. 367), and by this I think Scanlon wants to say that *commonsensically* we think that this deficiency includes "a failure to be moved by certain considerations that we regard as reasons" (ibid.).

Thus, the idea is that commonsensically we think like objectivists, who claim that we have a reason to do things even though that reason will not reflect our preexisting motivations. This is precisely because, according to this line of thought, commonsensically it would seem that all the things that we can say to the abuser indicate that he has a reason to stop, regardless of his current motivations.

John Searle (2001) gives another example that aims to expose the unintuitiveness of Williams' subjectivism. The passage is worth citing in full:

> Suppose you go into a bar and order a beer. The waiter brings the beer and you drink it. Then the waiter brings you the bill and you say to him, 'I have looked at my motivational set and I find no internal reason for paying for this beer. None at all. Ordering and drinking the beer is one thing, finding something in my motivational set is something else. The two are logically independent. Paying for the beer is not something I desire for its own sake, nor is it a means to an end or constitutive of some end that is represented in my motivational set. I have read Professor Williams, and I have also read Hume on this subject, and I looked carefully at my motivational set, and I cannot find any desire there to pay this bill! I just can't! And therefore, according to all the standard accounts of reasoning, I have no reason whatever to pay for this beer. It is not just that I don't have a strong enough reason, or that I

> have other conflicting reasons, but I have zero reason. I looked at my motivational set, I went through the entire inventory, and I found no desire that would lead by a sound deliberative route to the action of my paying for the beer.' *(Searle, 2001, p. 27)*

Searle (2001, p. 187) thinks that by ordering a beer in a bar we *create* a reason to pay for it, a reason that is external to the agent's prior motivational set. In so far as Searle's example shows that we have external, subject-independent reasons for action it nicely dovetails with Enoch's (2005) view that the commonsensical conception of reasons favors non-subjectivist theories of reasons and the objectivist commitments of our normative discourse.[31]

I will not try to refute the claim that the commonsensical conception of normative reasons is committed to the existence of robust normative facts as encapsulated in objectivist theories. However, one could argue that common sense needs revision. This line of thought leads us to Enoch's fourth possible answer that a subjectivist might offer to the question of why we should idealize. If we grant that our actual practice, as well as discourse related to reasons and normativity more generally, are committed to realist/objectivist presuppositions then the 'natural' answer to the question of why one would want to be a subjectivist idealizer is because one thinks that there is something wrong with commonsensical presuppositions. And therefore the revisionary account that includes idealization should be viewed as "the best reconstruction of the relevant discourse" (Enoch, 2005, p. 785). I will discuss this point more directly in the next chapter.

In what follows, I will pursue the issue from a different angle. As mentioned at the beginning of this chapter, I will argue that at least subjectivist accounts that have response-dependentist flavor can provide a 'natural answer' to the question of why we should idealize. I will argue this by drawing from arguments about color perception. Thus, in the next section I will start with the case of color perception, and will then draw some lessons for how a subjectivist about normative reasons can similarly provide a 'natural answer' to the question of why we should idealize.

[31] However, I must add that, given that Searle (2001, pp. 170-171) wants to give a naturalistic account of reasons, the fact that he thinks reasons are relational (ibid., see chapter 4) and can be *created* (ibid., pp. 186-187) is not very compatible with objectivist views about normative reasons as expounded by Enoch (2011) and Parfit (2011a; 2011b), for example.

3.3 Response-dependence, colors, and 'the natural answer'

I will start by noting that I do not find Enoch's (2005) cases of idealization that have a clear explanation in terms of tracking perceiver-independent facts very persuasive or plausible. The only thing that is necessary for an account to legitimately introduce some kind of idealization is the account's potential to sustain the difference between *appearance* and *reality.* And this is what response-dependentist accounts are able to provide.

The case with which I want to draw an analogy concerns color perception.[32] Making this analogy will enable me to do two things. First, it will show how a possible subject-based account in the non-normative domain can sustain the distinction between veridical perception and 'what seems to be the case'. Second, it will enable me to show that even if one has a revisionary account of some domain one can still provide a 'natural answer' in order to justify idealization.[33] That is, Enoch claims that even if one can provide a revisionary account of reasons (and reasons for the revision) nevertheless the "revisionist idealizer cannot rely on the natural answer [to justify the idealization] any more than the non-revisionist idealizer can" (Enoch, 2005, p. 786). The following short discussion of colors will show the falsity of the latter statement by showing that Enoch construes the applicability of the 'natural answer' too narrowly.

Color, phenomenologically, is to a normal observer[34] presented as an intrinsic property of material objects (Clark, 2000, pp. 13-14; Giere, 2006, p. 25). However, empirical research on color perception provides good reasons for thinking that colors are not just intrinsic, objective properties of external objects (cf. Palmer, 2000, p. 95).[35]

---

[32] This analogy and the more general analogy between *secondary properties* and normative properties is well established in the contemporary philosophical literature (McDowell, 1985; Miščević, 2006; Wiggins, 1987).

[33] At this point, a caveat is in order. There are some philosophers who argue that data on color perception could plausibly be interpreted in a way that is compatible with realism and objectivism about colors (see e.g. Byrne & Hilbert, 2003; Tye, 2000). Even though there is good evidence for thinking that colors are not wholly objective and intrinsic properties of objects, some of which I provide in the main text, I do not want to commit myself to any strong conclusion about the ontology of colors. For the purposes of the analogy, it is enough to say that there is a respectable view according to which colors are response-dependent properties. And according to this view, a distinction can be made between veridicality and appearance, to which an idealization condition can be applied.

[34] By 'normal observer' I mean a person whose visual system functions normally in the statistically average sense of the word.

[35] In framing the present section on color perception I rely on Giere's (2006, chapter 2) summary of the scientific evidence indicating that color cannot be completely objective or an intrinsic property of material objects.

There are at least two related reasons for this conclusion. First is the structure of the hue-dimension of the color space. The human visual system differentiates four-color hues: red, green, blue, and yellow. What is important about the hue of the color space is that structurally it has a circular form.[36] If colors could be reduced to some physical properties then the natural reductive base would be the physical properties of the spectral reflectance of a surface and the wavelengths of light that get reflected from those surfaces (Giere, 2006, p. 26; see also Hardin, 2003). If that were the case then we would be able to say, for example, that object O is red because it reflects the light of the wavelength X. However, any "simple identification of perceived hues with single wavelengths is" not possible because the structure of the wavelength is linear as opposed to the circular structure of the hue dimension of the color perception (cf. Giere, 2006, p. 18; see also Palmer, 2000, pp. 98-99).

The second reason is closely connected to the first and it involves the phenomenon called *metamerism.* Metamerism refers to a phenomenon in which light of *different* wavelengths can produce the *same* phenomenal color experience in a normal observer of surfaces (cf. Giere, 2006, p. 21).[37] Moreover, across different individuals different color experiences can be produced by the same combination of wavelengths presented in the stimuli (Clark, 2000, p. 11). However, even though the experience of a particular color cannot be identified across individuals with one particular class of naturally identified physical stimuli, the structure of the color space remains the same for all normal observers. Thus, for every normal observer we can construct a model of a color quality space in which red-green and blue-yellow will be opposed to each other and the identity of a particular color will be identified by its place in the color space.[38] Because of the circular form of the hue of the colors, for every normal observer green will be defined negatively in relation to yellow and blue. That is, it can be defined as not being yellowish and not being bluish. Other colors, such as orange, for example, would be defined in positive relation to other colors, that is, for every normal observer orange will be identified as a color that is in between red and yellow. What is important here is that even though different

[36] In geometric terms, red–green hues form one continuous axis and blue–yellow the other. Together they form a hue circle (cf. Giere, 2006, pp. 17-18; see also Palmer, 2000, pp. 98-99).

[37] All combinations of wavelengths that have the same impact on the visual system are called *metamers* (cf. Clark, 2000, p. 6). Metamerism is explained in terms of the opponent process theory of color perception that some authors refer to as the standard model of color perception (ibid., p. 10; see also Palmer, 2000, chapter 3).

[38] According to the opponent process theory opponent colors such as red–green, blue–yellow, and black–white oppose each other in the sense that they cancel each other out so that in normal circumstances there will be no combination of stimuli that produce the experience of greenish red or bluish yellow color.

stimuli can produce different color experiences in different subjects, once we identify which combination of wavelengths produce which color experience in a certain subject then we are in a position to infer the color quality space of that person. That quality space will, in the structurally relative sense, be the same across *normal* individuals that are members of the same species (cf. Clark, 2000, pp. 11-12).

So far I have been mentioning only the purported identification of color experience with physical stimuli that impinge on the visual system. Someone could object that colors could be identified with dispositions of object surfaces to reflect light of a certain wavelength, their so-called spectral-reflectance surfaces (Giere, 2006, p. 26; Hardin, 2003). Nevertheless, this *objectification* of colors to mind-independent properties of objects cannot work. The reason again is the phenomenon of metamerism. Because of the metamers, surfaces that have different spectral reflectances can produce the same color experience and therefore the only thing that groups those surfaces is the fact that they produce the same color experience, not their physical properties (cf. Giere, 2006, p. 26).

Even though it is plausible to think that scientific evidence points to the view that colors are not intrinsic properties of objects (Giere, 2006, chapter 2; Hardin, 2003; Palmer, 2000), the scientific value of the concept is not lost, and neither is its utility in everyday life. For example Giere (2006, p. 31) writes that color concepts have a role in evolutionary explanations of the selective advantages of visual systems that perceive color. It is supposed that color vision, for the most part, evolved for object recognition and identification. Color enables organisms to identify objects as conspecifics, potential mates, edible, etc. (Mollon, 1989). In particular it is often hypothesized that trichromacy in primates evolved as an adaptation for finding ripe fruit or edible leaves (Surridge, Osorio, & Mundy, 2003). So the ability to recognize a fruit, for example, by its color, in normal conditions where color is a cue for its ripeness, might have had, and typically does have, a long term consequence for the primates' life and fitness in general. This provides with a reason to distinguish the 'true' color of the object from its unreliable appearances in different light or sensory conditions.

Another reason for distinguishing the veridical perception of colors from appearances comes from the phenomenon of color constancy.[39] We intuitively "distinguish between color *appearance* and color *reality*. […] [W]e think we know someone's red BMW is really red even

[39] Color constancy refers to "the stability of perceived object color across changes in viewing conditions" (Wright, 2013). So for example we perceive an apple as red whether the apple is placed in deep shade or is illuminated by white sunlight.

though it does not appear red at night in a parking lot illuminated by sodium vapor lamps" (Giere, 2006, p. 24). From this point of view, Giere (ibid.) concludes that "[c]olor seems no different from other physical properties, like shape." Nevertheless, since color concepts definitively refer to objective phenomena to some degree,[40] and are useful in everyday life, a natural move is to revise our color discourse and try to provide truth conditions that can accommodate the evidence that seems to favor the view that colors are actually mind-dependent properties.[41] According to one influential line of thought, response-dependence theories of secondary qualities provide just such a minimal revision that preserves the objectivity (or intersubjective validity) of colors and at the same time incorporates a subjective element in their truth-conditions (see e.g. Giere, 2006, pp. 31-33; Miščević, 2011; 2012).

The main point I want to make here is that even though colors, in light of the relevant evidence, cannot be construed as mind-independent properties of objects we can still provide a 'natural answer' to the question of why we would want to idealize when we judge whether an object is of color X or Y. Pace Enoch (2005), a revised discourse about some domain can retain its distinction between *appearance* and *reality* even though the facts in question do not count as mind-independent. We normally say that if object O looks red to me then I am justified in believing that it is *true* that object O is red. However, if my reliable friend points out to me that object O is not red, but that it just looks red to me "because of peculiar lightning conditions" (Pollock, 1987, p. 484) then I will recognize that my reason for thinking that O is red is defeated. That is why it is natural for response-dependentist theories of colors to introduce idealizing conditions in their accounts (see e.g. Miščević, 2011).

To give a very simplistic example we can say that object O is red when O looks red to agent A and A is not in defeating circumstances (such as looking at O under peculiar lightning conditions, or when drugged and/or hallucinating, etc.). Therefore, the conclusion of this section is that Enoch's (2005) claim that only mind-independent realists about a particular

[40] This point is emphasized in the following quote: "In our perception of object color all these elements are involved; there is light radiation, which is selectively absorbed and reflected in different ways by objects that differ physically and chemically; when the light rays coming from objects are imaged on the retina, they set off a complex series of neural events that are associated with the visual experience of color" (Hurvich 1981, 52, as quoted in Giere, 2006, p. 31).

[41] Here I do not mean to imply that our discourse about color really needs revision, because the idea that colors are not mind-independent properties has been prominent at least from Galileo onwards. And so I suppose that for some people, not only vision scientists, the idea that colors are not mind-independent could be a part of their commonsensical view of things.

domain can provide a 'natural answer' to why we should idealize is not correct. The case of color perception provides a paradigmatic example in which we have potentially established that some property is not mind-independent[42] but where the 'natural answer' to why we should idealize still applies. Moreover, it could be seen as justifying the intuition that there is still some difference between how things *seem* to be and how they really *are*, namely a difference that can be discerned from a better epistemic position.

3.4 Response-dependence, reasons, and rationality

In this section I will try to show how the analogy between colors and reasons holds and in what sense reasons could be construed as response-dependent. I will start by offering some ideas about the structure of reasons.

There is one way in which the structure of reasons can be construed as relatively non-normative and therefore uncontroversial from the naturalistic perspective. This can be seen in the following line of thought. It is often said that reasons are facts that count in favor of something. Some authors (e.g. Parfit, 2011a; Scanlon, 1998) take this claim to represent an irreducible normative property of facts that provide reasons. However, I contend that saying that reasons are facts that count in favor of something is to a certain extent a metaphor that indicates an anthropomorphic element in our idea of a normative reason. When we try to make a conscious decision we often represent the situation as involving a list of pros and cons. In this context, the idea of counting in favor makes sense because the deliberator is a person who does the counting. However, in some more objective (agent independent) sense it is not clear how facts by themselves could count in favor of anything, as if they had some kind of intrinsic agency. Someone could say that this in some way already shows that reasons are to a certain extent response or mind-dependent. But that is not my line of argument at this stage. Here I only want to say that when the 'counting' part is disregarded what is left is the idea that facts (that are reasons) *favor* something (e.g. adopting an attitude, etc.). However, my claim is that there is an interpretation of the notion of *favoring* that does not imply the existence of an *irreducible* normative property of the fact in question. This interpretation will show the objective side of reasons that does not imply intrinsic normativity.

[42] I am aware that this claim is controversial since there are some philosophers who still want to hold on to the claim that colors are objective, physical properties of objects (see Hardin, 2003).

There is nothing *intrinsically* normative about saying that some fact favors something. For example, we normally say that natural selection favors the occurrence of some trait. Alternatively, we say that the evidence favors some hypothesis, where this means that the evidence indicates the probability of the hypothesis being true (cf. Parfit, 2011b, pp. 504-506). Similarly, we can say that reasons as facts that *favor* something in the latter sense have the following structure. In a perfectly normal way, we can say that reasons favor some attitude, for example, if the adoption of that attitude will serve some purpose or be conducive to accomplishing some goal. For example, if someone has the goal of learning the fundamental physical structure of the world then that person has a reason to believe in the Standard Model of particle physics. The reason is grounded in the fact that there is strong experimental and theoretical evidence that the model accurately describes the world of quantum phenomena. In this sense experimental and theoretical considerations provide reasons or grounds on the basis of which one forms (or can form) beliefs about the fundamental structure of the world. Similarly, the ability to use practical reason exposes the structure of favoring, at the most basic level, as an instance of discerning and using means–ends relations that are grounded in the natural world. For example, we can say that the fact that there is milk in the nearby grocery store favors the formation of an intention to go there, because that is the efficient way to come into possession of a box of milk.

The considerations just adduced are nicely captured in Stephen Finlay's *end-relational* theory of normativity (Finlay, 2006; 2009; 2010a). While Finlay proposes his theory as a reductive analysis of normative reasons and normativity in general, in the present discussion the theory is significant because it exposes some reference points for comparing the response-dependentist theory of reasons to the response-dependentist theory of color vision.

According to the end-relational theory of reasons, "a fact is a reason for Φ-ing, relative to a system of ends E, iff it explains why Φ-ing is conducive to E" (Finlay, 2006, p. 8). Thus the structure of the *counting in favor of* relation is explained in terms of the relation between the end or purpose of some act or attitude and the means that are *conducive* to furthering that act or attitude (cf. ibid.). The "explanation why something is conducive to goal" can be thought of here as nothing other than what *makes* it the case that Φ is conducive to E. For example, the fact that the tire is flat makes it the case that pumping the tire will promote the goal of driving home.[43]

[43] This reading of the word 'explanation' is based on Broome's (2013) construal of reasons as normative explanations. Roughly, Broome writes that normative reasons explain why something ought to be the case in the

This view enables us to extend Skorupski's (2010) description of the structure of the reason-relation introduced in the first chapter of this dissertation (see subsection 2.3.6). Now we can say that the non-normative part of the *counting in favor of* relation concerns the relation between four basic ingredients: the fact that *p* is a reason for *A* to do *Φ* in relation to an end *E*. And p is a reason for A to do Φ because it explains why Φ-ing is a good means of accomplishing E.

The end-relational theory of reasons explains the objectivistic grounding of reasons and in what sense externalist intuitions about reasons seem to be true. In general, facts provide normative reasons in relation to a system of ends. This is how Finlay frames the point:

> The end of prudence is personal wellbeing, so a fact is a prudential reason for me to Φ iff it explains why Φ-ing is conducive to my wellbeing. Epistemic ends include having beliefs that are true, so a fact is an epistemic reason for believing that p if it explains why p is likely to be true. The ends of playing chess are checkmating and avoiding being checkmated (within the rules), so a fact is a chess reason for move M iff it explains why M advances these ends. There are as many kinds of normative reasons, therefore, as there are different systems of ends: there are reasons of faith, revenge, love, lust, and football, there are legal, medical, military, artistic, political, and scientific reasons. *(Finlay, 2006, p. 9)*

The objective part of the favoring relation that is grounded in means–ends relations can be compared to the objective properties of color vision. In order to see colors certain combinations of light wavelengths need to be reflected from an object with certain reflectance properties and then impinge on the eyes' retina. This objectivity grounded in means–ends relations explains how we might be wrong about our reasons in at least one way. Namely, if I do not know that Φ-ing is conducive to my end E than I might wrongly evaluate what my reasons are. For example, I might think that eating mud will keep me healthy. But since this belief is wrong I do not really have a reason to eat mud, since it would not further my goal of being healthy. In this sense, we have a natural explanation (cf. Enoch, 2005) of why we want to take idealization into account when we are trying to determine our reasons for Φ-ing.

---

sense that reasons make ought-statements true (ibid., p. 48). He draws an analogy with natural selection: natural selection explains evolutionary change in the sense that it makes evolutionary change occur. The textual evidence that Finlay has in mind for this sort of meaning of the word 'explanation' is provided in (Finlay, 2006, p. 9, footnote 17). There he states that he revised the definition of a normative reason because in the first version the definiens, which said that reasons *indicate* that Φ-ing is conducive to E, blurred the distinction "between reasons for Φ-ing and reasons for believing that one ought to Φ."

The second component of the favoring relation concerns the ends themselves. The ends or system of ends might be of different varieties, and we do not want to say that they are of equal *importance* and that they provide equally strong reasons (Goldman, 2010; Parfit, 2011a; Scanlon, 1998). For example, most of us would agree that reasons stemming from moral ends, such as the end of sustaining cooperation on fair terms (Rawls, 1971), have priority over those reasons stemming from antisocial goals, such as being the best serial killer in history. Furthermore, this simple objectivist view of reasons seems to multiply reasons excessively; every goal that one might adopt or every desire from which a goal might stem could provide a normative reason for something (cf. Finlay, 2006, p. 9; Schroeder, 2007, chapter 5).

A starting point for a resolution of this problem is the introduction of a distinction. We can make a distinction between reasons there *are* and reasons that we *have* or that apply to us in so far as we are potentially responsive to them.[44] The reasons that there are correspond in general to the objective dimension of reasons as discussed above. The reasons that we have might be identified with reasons that stem from ends that we especially *care* about.

Finlay (2006; 2010b) makes a distinction between normative reasons as explanations of why something is conducive to some end and reasons that *matter*, i.e. reasons that stem from ends that matter to us. So reasons that matter would constitute a narrower class of normative reasons, namely those that follow from ends that matter to us and reflect our concerns and what we care about (Finlay, 2006, pp. 17-19; see also Frankfurt, 1988a for the importance of the concept of caring in our normative thinking).

Reasons that we care about limit the scope of normative reasons because our concerns for certain classes of ends limits the facts that we will treat as reasons. In addition, the priority relations between our ends will impose constraints on the relative priority of the ensuing reasons and their strengths. For example, at least nominally philosophers emphasize moral reasons as a kind of reasons that have possibly overriding authority over us (Brink, 1997; Joyce, 2006). If this is true about moral reasons, then we can say that these philosophers have recognized and thusly delineated the class of reasons that matter to us and which structure our thinking. In this

[44] For example, Gaus (2011, p. 233) and Goldman (2010, chapter 2) stress the importance of making the distinction between reasons that there are and reasons that we have. In recent years the notion of 'reasons that we have' has been put under philosophical scrutiny (Schroeder, 2008). However, for my purposes it is enough to grasp the intuitive distinction between reasons that are there because there are facts that are conducive to some ends and reasons that one might have because of the ends that one cares about.

sense, it is supposed that the saliency[45] of the facts that are moral reasons has the capacity to override, diminish, or even nullify the propensity of some facts to act as reasons. For example, my capacity to care about the well-being and rights of other human beings normally disables me from even considering the alternative, in which people's basic rights are undermined as a potential reason for action, even though such an action according to some set of ends might be supported by some other reason.

I want to argue that the ends to which reason-claims are relativized represent the subjective part of the 'counting in favor of' relation. In this respect, reasons are also similar to colors, construed as response-dependent properties. To use Euthyphro's dilemma once again; some ends matter and provide stronger, overriding, etc. constraints on attitudes and actions because we care about them; it is not the case that we care about them because they have some intrinsic, reason-giving, irreducible normative property (cf. Finlay, 2006, p. 17). Of course, things might be such that we judge them valuable and because of that judgment we start to care about them. However, on the present view the simple explanation for this fact is that when we judge something to be valuable we judge it so because it stands in a means–end relation to something that we ultimately care about intrinsically. For example, most people care about the wellbeing of their children intrinsically. We do not normally search for reasons to adopt the goal of caring about our children, rather this intrinsic concern constrains the rest of our goals and makes salient facts that favor actions and attitudes that sustain and accomplish that goal (cf. Finlay, 2006, p. 17; Goldman, 2010, p. 9).

3.4.1 Personality traits and rationality: VM patients and the successful psychopaths

Why should we think that reasons have this response-dependent dimension? In the case of color vision we saw that the strongest argument for the dispositionalist view comes from the phenomenon of metamerism. Colors cannot be identified with light's wavelengths because different wavelengths can produce the same color experience. I think that one can similarly argue, although on different empirical and conceptual grounds, that judgments about what we

[45] A number of authors use the concept of saliency in a similar way. My concept of saliency in this context is similar to that provided by Schroeder (2007, section 8.3, 8.4). When an agent has a certain desire then a number of considerations become salient to the agent, namely, those that would explain why performing certain actions would satisfy that desire (cf. Schroeder, 2007, pp. 156-157). In the terminology that I use here, following Finlay (2006), this amounts to saying that the goals that intrinsically matter to us pick out those facts as reasons that serve as explanations for why performing certain action(s) would be conducive to satisfying those ends.

care about most do not solely come from the objects of our judgments but also reflect our affective and personal traits.

The evidence for this view comes from social and moral psychology. There is a great deal of evidence showing that people's attitudes about practical affairs are to a large extent based on emotional responses to those affairs (for a review see Haidt, 2007). In their seminal study Haidt, Koller, and Dias (1993) showed that when people are confronted with scenarios that elicit strong emotional reactions they persevere in their evaluations even though they cannot provide justifications in support of their judgments. This phenomenon was later named 'moral dumbfounding' (Haidt, 2001). Later this phenomenon was more explicitly investigated. For example, participants were asked whether it is ok for a brother and a sister to have consensual sex. Subjects answered negatively, and then the experimenter asked them to justify their judgments. In the process, the experimenter played devil's advocate and pointed out the flaws in the participants' justification. For example, if the participant said that the problem was that the sister might get pregnant the experimenter would point out that in the example it is stipulated that the siblings will have safe sex, etc. What is interesting in this study is that a significant number of people did not change their initial judgments even though they could not find any convincing reason for upholding their judgments.

From the perspective of today's knowledge this is not surprising, because we now know that normal human judgments about practical affairs are heavily influenced by emotional processes. This is also confirmed by neuropsychological data. For example, when people evaluate moral dilemmas such as the trolley problem or when they make economic decisions brain studies have shown that the normal pattern of brain activation includes activity of brain areas associated with processing and regulation of emotional stimuli (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003).

Furthermore, studies have shown that when brain areas associated with processing and regulation of emotional states, such as the orbitofrontal and ventromedial prefrontal cortex, are damaged people stop caring about the things they normally care about. For example, patients with lesions in the ventromedial prefrontal cortex (VM patients for short), who used to be responsible members of their communities, stop caring about interpersonal relationships and about their families. They lose interest in their jobs and become irresponsible. They tend to be antisocial and develop an abnormal pattern of reactions to moral problems and dilemmas (Koenigs, et al., 2007). This occurs even though the capacities underlying their general cognitive processes and general intelligence remain intact (Damasio, 1994).

From these data, Antonio Damasio (ibid.) draws general conclusions about the importance of the emotions and affective processing of stimuli for rational behavior. However, what is important for the present discussion is that the capacity for normal processing of emotions underlies our ability to *care* about certain things and to be motivated by those things. In the language of reasons, we can say that those patients become unresponsive to certain kinds of reasons. And the natural explanation for this unresponsiveness is that they stop caring about certain things. For example, their marriages fail because they tend to lack concern for their spouses, which in turn makes them oblivious to reasons for being kind, helpful, and loving, reasons that would be otherwise salient to a person who cares about her marriage.

It could be objected that VM patients do not represent a good case for the claim that our reason-grounding ends are determined by affective states that underpin what we care about because VM patients might be less rational than other people without similar problems. By rationality I mean here the very narrow notion of avoiding self-defeating behavior or patterns of thought (see Goldman, 2010, chapter 2, section III). On this view, for example, it is irrational to adopt something as a goal but not to adopt the necessary means for accomplishing that goal, because having a goal involves intending to accomplish it and not intending to adopt the necessary means includes not intending to accomplish the goal, which is self-defeating. On this view of irrationality, it could be argued that VM patients are irrational. Namely, research on gambling tasks has shown that VM patients make more disadvantageous decisions than control groups when their proclaimed goal is to make economically advantageous choices (Bechara A. , Damasio, Tranel, & Damasio, 1999; Rolls, Hornak, Wade, & McGrath, 1994). This finding could be generalized (Damasio, 1994) so that we could say that they do not care about certain ends because they are irrational and therefore unresponsive to reasons that actually apply to them (cf. Maibom, 2005).

This objection points to something significant, namely the importance of rationality[46] in assessing the ends that determine the reasons that we have (Gaus, 2011; Goldman, 2010; Smith, 1994; Williams, 1981). However, it does not change the general lesson about the subject-dependence of the importance of the goals that we deem valuable.

There are other, less controversial cases of people who do not have rational problems in the minimal sense of rationality but do have *shifted* emotional responses to the well-being of other people and to social values in general. These are so-called successful psychopaths (Hall

[46] In the minimal sense of avoiding self-defeating actions, intentions, and beliefs. However, see chapter 6 of this thesis for a more thorough discussion of the rationality of psychopaths.

& Benning, 2006). The term 'successful psychopaths' delineates a group of people that have affective and interpersonal deficits, and from a moral point of view possess problematic traits; they are manipulative, cunning, glib, guiltless, remorseless, lack empathy, etc., and possess other psychopathic personality traits (ibid.). However, unlike so-called *unsuccessful psychopaths*, they do not have police records, and/or, for the most part, they manage to stay under the police radar, even though they are prone to engage in immoral and illegal activities (Glenn & Raine, 2014, p. 149).[47]

Unlike unsuccessful psychopaths and VM patients, studies show that successful psychopaths do not show abnormalities in the brain regions that underpin executive functions (Gao & Raine, 2010). For example, studies have shown that successful psychopaths may have deficits, albeit to a lesser degree than unsuccessful psychopaths, in the orbitofrontal cortex and amygdala, which are the brain areas that underlie the representation of the affective value of stimuli (Glenn & Raine, 2014, pp. 151, 156). However, successful psychopaths exhibit normal or superior functioning of executive functions (Glenn & Raine, 2014, p. 153). This is also supported by neuroimaging studies indicating that the brain areas that underpin cognitive functioning, such as "the superior parts of the frontal lobe, the parietal lobe, and the anterior and posterior cingulate", show normal or even enhanced activation in these individuals (ibid. p. 156). It is speculated that the normal or enhanced functioning of executive functions (in combination with tendencies to engage in immoral action) plays a 'protective' role for successful psychopaths and decreases their chances of being detected and processed by society's law enforcers (ibid. p. 151; Sifferd & Hirstein, 2013, p. 135).

Thus, from this perspective we can conclude that as far as the brain correlates that underpin the executive functions are concerned, successful psychopaths are rational and thus a difference in rationality cannot explain their behavioral differences. What can explain the differences are the personality traits that they share with the rest of the psychopathic population; these include the above-mentioned affective and interpersonal deficits. Translated into the idiom that I am currently using, the differences lie in the concerns and ends that they care about, which differ from those cared about by the population that score low on PCL-R.

[47] I use the example of successful psychopaths because there is an emerging consensus in the literature that they are the sort of people who are rational but predictably immoral (Aaltola, 2014; Ramirez, 2015; Sifferd & Hirstein, 2013), which can indicate which reasons and constraints are salient to them as functions of the ends they care about.

Since the differences between successful psychopaths and non-psychopaths are currently best explained by differences in their affective responses and emotional processing (Glenn & Raine, 2014, pp. 157-158), we have reason to think that judgments about what matters and what is worth caring about are underpinned by response-dependent properties. The argument for this conclusion is based on the consideration that affective responses underlie which ends we will find valuable and therefore which reasons will be salient to us as intrinsically important. Furthermore, this consideration is grounded in the fact that rational capacities cannot explain the differences in attitudes and behavior that an important class of psychopaths exhibit in their normal functioning.[48]

Now we can complete the analogy with the case of color vision. Just as we need to have a certain sort of perceptual apparatus (adjusted for trichromacy) to perceive objects as colored, similarly we need to have a certain structure of caring in order to *perceive* certain facts and ends as counting in favor of something or providing reasons. Psychopaths seem to lack empathy, which, for example, makes them in many circumstances oblivious to other people's feelings and as sources of reasons to be treated with respect. In this respect, they seem to be blind to certain reasons in the same way as the color-blind are blind to colors.

It is important to recognize the limits of the analogy. In certain respects, the analogy between colors and reasons stops here, but in others it goes deeper. The analogy breaks here because capacities fix truth conditions differently in the case of color and reason attributions. For example, while it seems okay to say that agents with normal visual capacities fix truth-conditions of statements about colors, it does not seem right to say that agents with normal affective capacities fix what there is reason to do in general. The difference seems to lie in the fact that normative reasons *apply* to rational agents in virtue of their capacities, preferences, beliefs, etc., while colors do not *apply* in a similar way to anything.[49] On the other hand, the

[48] Moreover, Osumi, and Ohira (2010) hypothesize that emotional detachment enables people with heightened psychopathic personality traits to reach more rational decisions. They asked college students to play the ultimatum game and found that people with a high tendency to psychopathy made more economically rational decisions by accepting more unfair offers in the game. The explanation for this is that normally people have an emotional aversion to unfair offers. Since people with heightened psychopathic traits have diminished emotional responses they are not bounded by them like nonpsychopaths and therefore have the ability to reach more optimal decisions, at least in contexts of economic exchange.

[49] Indeed, the talk about colors applying to certain agents rather than others, in the same way that we say that reasons apply differently to different people, does not make much sense. Compare the following two sentences. The reason to take your child to an amusement park does not apply to you since you do not have children. The fact

analogy between the capacities that ground color vision and reasons runs even deeper, in the sense that there is nothing *a priori* irrational about lacking certain affective capacities, in the same way that there is nothing essentially irrational about being color-blind. Thus, to circumscribe the analogy, we can say that given their perceptual capacities color-based considerations will not normally provide color-blind people with reasons for action, in the same way that affect-based considerations will not normally provide practical reasons for people who do not structurally have the capacity to care about them.

3.5 Why idealize? Deeper concerns, competing desires and non-parametric decisions

The question with which I opened this chapter was: why idealize if we think that reasons have a subjective or response-dependent aspect? Earlier I responded that one reason for idealizing relates to the objective part of the *counting in favor of* relation: namely, given our ends, preferences, concerns, etc. we want to know which facts are conducive to satisfying them. This is something about which we can be wrong and therefore it often pays to search for better epistemic vantage points.

The second important reason for idealizing relates to fixing the standards of evaluation of our ends. Our cognitive abilities mean that we are creatures that not only discuss and evaluate the means to our ends but also the goals and ends that determine the reason-giving power of the relevant means. The justification is functional. We are creatures that have preferences, desires, ends, concerns, etc. that are often incompatible and compete for our cognitive resources. Bluntly speaking, the function of a desire to Φ is to make you Φ. However, not every desire will reflect your deeper concerns (Goldman, 2010), the concerns that make up your identity, for example. Therefore, when we reflect on whether to satisfy some desire we are trying to find out whether satisfying that desire would be worthwhile given the structure of our concerns that constitute us as the persons that we are. In addition, our concerns might be something that we can be wrong about or that we need to try to discover.

Furthermore, when we have a desire for Φ and a desire for Ψ that are incompatible, this presents us with a practical conflict that we might need to solve in order to act successfully. In the words of Harry Frankfurt (1988b), competing desires, preferences, or ends provide a context in which we need to take a *stance* as to which desire, preference, or end we are going to make

---

that this object is green does not apply to you since you do not see colors. While the former sentence is completely clear and informative; the latter is not. Furthermore, it is not straightforwardly clear how we should interpret the latter sentence.

our *own*, or with which we are going to *identify*. This might be construed as a question of value for which reasons need to be provided. However, here I am construing the situation naturalistically, that is, assuming that we *are* the sort of creatures for which these kind of conflicting situations are reality and that it is in our nature to solve them by using our capacities for reflection and deliberation about our options. To solve these kinds of situations of conflicts it is often effective to have at least a minimally transitive ordering of desires, preferences, or ends and a settled set of goals that provide consistent constraints for which other goals might serve as reasons for action.[50]

Any attempt to reach a consistent set of desires will usually include hypothetical thinking (idealizing), because our natural instincts and inclinations lead to conflicts between desires. For example, we desire to stay healthy and desire to smoke; we desire to procrastinate and desire to work; we desire what is present over bigger rewards in future; we desire to constrain ourselves and to indulge in the delights of the present moment, etc. (cf. Ainslie, 2001). These conflicts make it hard for us to successfully act in the world in a way that reflects the persons we are or the persons we want to be. The function of idealization in this context is to give one a vantage point from which the solution corresponding to one's deep concerns can be sought and implemented.

The notion of a defective desire has a place in this framework. These would be those desires that cannot be justified from the perspective of a person's deep or intrinsic concerns (Goldman, 2010). The reason for giving priority to our deep concerns is that they constitute our identities and what makes things important to us (Frankfurt, 1988a). Thus, we might say that in this respect our deepest concerns and the desires that stem from them have a default evaluative authority. Of course, their being default implies that they can also be revised under pressure from, for example, experience or deliberation, but still their importance derives from the fact that they constitute the agent's stance (Frankfurt, 1988b) or the point of view that grounds the deliberation. In this sense, if a desire to turn on radios or eat saucers of mud does not reflect a person's deep concerns she will not have a reason to do those things, because they would not be important for her and on reflection such desires would probably be revised or repelled.

Other salient ways in which desires can be defective is if they are self-defeating or dependent on other factors that lead to self-defeating action, such as errors in rationality

[50] See Frankfurt's (1988b, pp. 170-171) discussion of the importance of the operations of *integration* (making an order) and *separation* (providing constraints on which desire is admissible for acting upon) for solving conflicts between competing desires.

(including errors in the inputs to reasoning processes). For example, if a person has a desire to p and not to p then this desire will probably be experienced as defective and the person will try to get rid of it or revise it because it cannot be satisfied. In addition, if the desire to eat mud depends on the belief that by doing this regularly one will keep one's children healthy then such a desire could also be deemed defective.

However, even though on the present view of reasons we can make sense of the need for idealization and the idea of defective desires, it does not give us a principled way to distinguish those desires that are defective or irrational in some respect from those that are not. Therefore, unlike some authors, I do not see any *a priori* reason for saying that a person who wants to devote her life to counting blades of grass is irrational or that she has defective desires. On the contrary, in the light of how Rawls (1971) construes the case, that person seems very rational and able to secure the means for executing her life plan.

However, the present view on reasons allows for extensions that might enable us to put more constraints on what might typically count as a reasonable desire or the structure of concerns that will pick those reasons out. People are social beings that interact, cooperate, compete, etc., and these interactions put constraints on people's desires, beliefs, and structure of concerns. These considerations lead us to a third significant way in which idealization is important in determining what reasons a person has.

3.5.1 The interdependency of reasons and idealization

People do not only confront *intra*personal conflicts or problems that need to be solved but also *inter*personal conflicts or problems that stem from living together in diverse communities. In this context, the concepts of parametric and non-parametric decision situations are important. In parametric reasoning or decision-making, the parameters of the decision situation are set and do not change in response to a person's decision. For example, when a person is wondering whether to take an umbrella to work tomorrow it is rational for her to check the probability that it will rain tomorrow and reach a decision on the basis of this probability. However, when she reaches a decision the person does not need to worry further whether her decision will influence the weather prognosis, because whether it will rain or not is independent of her decision. The parameters of the situation are set before she makes a decision; she just tries to learn or guess the parameters in order to reach a satisfactory decision.

In non-parametric situations, the prospect of making a decision might change the parameters of the situation. For example, when playing rock, paper, scissors[51] Smith's winning strategy depends on predicting the choices that Jones will make. However, Jones' winning strategy depends on predicting what Smith will play. So their decisions are interdependent and affect each other.

These game-theoretic situations depict an important aspect of our reasons that is most salient in games of pure coordination (cf. Verbeek, 2008).[52] Suppose that my girlfriend and I plan to meet up but for this or that reason we have not decided where to meet. Furthermore, let us suppose that at this point we do not have any means of contacting each other. In such a case any meeting place we each choose will be as good as any other; the only important thing is for us to meet *somewhere*. In this situation it is clear that there is no independent reason for choosing one meeting place as opposed to some other. Our reasons would depend on the mutual anticipation of our choices and in effect our reasons would refer to each other. Verbeek nicely illustrates the situation:

> What reasons are there for my going to location 1? I have such a reason only if I believe that you will go to location 1. Why would I believe that? Well only if I believe that you believe that I will go to location 1. That is, only if I believe that you have a reason to go to location 1. What reason do I have for that belief? I have a reason for this belief, if I believe that you believe that I believe that you believe that I will go to location. In other words, I have such a reason if I believe that you have a reason to believe that I have a reason to go to location 1. My reasons for going to location 1 depend on your reasons for going to location 1 and vice versa. Our reasons are interdependent. *(Verbeek, 2008, p. 74)*

[51] Rock, paper, scissors is a game played with two or more players by simultaneously making hand gestures. A fist represents a rock, which beats scissors. The scissors gesture beats paper, which is represented by an open fist. Finally, paper beats rock because it can encompass it.

[52] A pure coordination problem refers to a set of games that have multiple Nash equilibrium points and therefore multiple, equally *good* solutions to the game. Nash equilibrium is one of the most important concepts in game theory. It refers to a situation in which all players are "simultaneously making a best reply to the strategy choices of the others" (Binmore, 2007, p. 14). So when the Nash equilibrium occurs no player has an incentive to unilaterally change her strategy, because the Nash equilibrium is a situation in which all players are doing the best they can. For example, the problem of deciding which side of the road to drive on is an instance of a coordination game. Driving on either side of the road is good enough as long as enough people are committed to driving on the same side. Moreover, no player has a reason to unilaterally change the side of the road on which she drives, because by avoiding coordination with others she would put herself (and others) in life-threatening danger.

This *interdependency* points to a significant aspect of reasons because it indicates another way in which idealization is important in developing the response-dependentist account of normative reasons.

To elaborate this point I offer a simple model, which in game theory is well known as the hawk–dove game. This game was originally employed by Maynard Smith (and other evolutionary biologists) for modeling interactions between organisms and their strategies that led to the evolution of cooperation (Binmore, 2007, p. 136). However, like many models in game theory, it can be used for structuring and theorizing about human interactions and, more generally, cultural evolution. The structure of the hawk–dove game is provided in **Figure 1**. The game is usually used to model situations where organisms compete for valuable resources. The terms "hawk" and "dove" are used to designate strategies that a player can use. Hawk is an aggressive strategy that always fights for resources when there is an opportunity. Dove is a more careful strategy; it only tries to attain resources when the competitor is another dove. If the competitor is a hawk, then the dove backs down. Since hawks always play aggressively then when they meet another hawk they are bound to fight. Since fighting itself is costly and nobody retreats, when a Hawk meets another hawk they both lose in terms of their payoffs.

		A	
		Dove	**Hawk**
B	**Dove**	**2** **2**	**4** **0**
	Hawk	**0** **4**	**-1** **-1**

Figure 1 (adapted from Binmore, 2007, p. 137)

When a dove plays against another dove than their payoff is equal. Everyone gets 2. If a dove plays against a hawk, then the hawk always wins. The hawk gets payoff 4 and the dove gets payoff 0. If a hawk plays against another hawk then they both lose, their payoff is -1.

Let us suppose that there are two agents (A and B) who find themselves in a situation that can be represented by **Figure 1**. For example, they need to decide who will get some valuable property.[53] How are they supposed to choose what to do? Since A and B are in symmetric situations they both have the same preference profiles. They both prefer to play hawk if the other is playing dove to playing dove against dove or hawk against hawk. They also prefer to play dove against dove than hawk against hawk.

The game represents a situation in which A and B's reasons are interdependent (cf. Verbeek, 2007, p. 247). If A decides to fight (play hawk) then B has a sufficient reason to retreat (play dove). If A decides to play dove then B has a sufficient reason to play hawk and *vice versa* if A is trying to respond to B's decisions. The problem with this situation lies precisely in the fact that A and B's reasons are interdependent. Since one's reasons for deciding depend on the reasons that the other agent has for deciding what to do, there is no rational way for them to decide what to do just on the basis of the reasons they actually possess.[54]

The hawk-dove model represents situations that do not seem to be so uncommon in real life (Binmore, 2007). However, if the game as presented here does not have a rational solution, how the problem is solved in real life? The theory of biological and cultural evolution provide us with an answer. The solution comes as a spontaneous and non-deliberate distribution of strategies in the population of organisms (including humans, in our case). For example, a certain proportion of the population of agents will some of the time play dove and some of the time play hawk, and during many rounds of encounters through selective processes an equilibrium between the proportion of individuals that play particular strategies will emerge (see e.g. Skyrms, 1996, chapter 1; Verbeek, 2007, pp. 147-148).

For example, one stable pattern of interaction that seems to solve the problem includes the following strategy: if a person finds a property (land, forms of energy, commodity, etc.) first

[53] We can suppose that payoff 2 means splitting the property, 4 taking the entire property for oneself, -1 not getting the property and, moreover, suffering injuries from fighting each other.

[54] Of course, we can always stipulate that A and B have some independent reasons for deciding to play one strategy over the other. For example, we could suppose that moral reasons count in favor of being a dove and splitting the property. However, if that were the case than the game would need to be construed differently, because the payoffs from **Figure 1** would not represent the import that moral reasons introduce. For example, playing dove would need to bring more payoff than playing hawk against a dove. However, this would miss the whole purpose of the introduction of the present hawk–dove model, because I want to show that situations in which reasons are interdependent could show why idealization is appropriate in response-dependentist accounts of normative reasons.

then she should defend it by fighting for it if someone refuses to grant her authority over the property. Therefore, the strategy would be that if you are first to come into possession of the property then you play the hawk strategy.[55] For this strategy to become stable, agents or organisms in the population need to be able to learn and change their strategies through encounters with each other. However, an ability for higher-order thinking or reasoning is not necessary for establishing this equilibrium of strategies.

Here we come to the main point of this section. If this strategy stabilizes in the population, then based on this pattern of behavior other individuals know what to expect and on the basis of that expectation they can reach decisions about how to act. For example, if the payoffs are set as in **Figure 1,** then on the basis of this recognized pattern of behavior A can reach a rational decision and decide to play dove when confronted with B, who was first to claim some property (resource, etc.). Moreover, based on the same pattern of behavior and on A's expectation that B will play hawk, B himself can reach a rational decision to play hawk. The reasons that A and B now have to decide have emerged from an established pattern of behavior and the expectations that those patterns ground.

There are two lessons I want to draw from this example. First, the reasons that A and B now have are response-dependent.[56] They depend on an established pattern of behavior, and since A and B are rational agents they also depend on A and B's higher-order expectations. That is, it is not just that A has a reason to give in because she knows that in this situation B will fight. The reason for giving in comes from A's expectation that B will fight because A believes that B expects her to give in. Similarly, B's reasons for fighting come from B's belief that A expects B to fight in this situation. A and B's capacity for rational, higher-order thinking enables the constitution of reasons for action that they otherwise would not have, and that is why we can say that this interdependency of reasons makes them response-dependent.

Second, idealization is important because it plays two roles. One is ontological: the capacity to think about what I would do (or what I would expect others to expect me to do) if I were rational, in the present category of situations, constitutes the reason for action that I have. The other is epistemological: in order to reach my reasons in this kind of situation I have to think about what I would do (or what I would expect people to expect me to do) if I were rational. Rationality is important in both roles because on the one hand it constitutes the

[55] It is important to note here that the strategy 'if you come second fight and if you come first give in' could also become a Nash equilibrium if enough individuals in the population were to conform to it.

[56] Alternatively, maybe we should say that they are *expectation-dependent.*

deliberative point of view[57] that in turn constitutes our reasons for action, and on the other hand it enables agents to track the reasons that they have in virtue of being rational, in the psychological sense of the word.

This perspective can help us to account for the intuitions that underpin Searle's example introduced in section 3.2. To repeat, in the example a person orders a beer in a bar and then refuses to pay for it because she does not have a desire to pay for it in her motivational set. If we model the situation as per **Figure 2**, we can explain where the problem lies. The established practice in our society is that when you order a beer in a bar you create an expectation to pay for it. Therefore, the interlocking set of expectations is that a customer A, by receiving a beer, expects that a bartender B will expect her to pay for the beer and for this reason will insist on getting the money for the beer (hawk strategy). Similarly, B will form the expectation that A will expect her to insist on paying and for that reason would be willing to pay for it (dove strategy).

The problem stems from the fact that if A decides not to pay then she will be violating these expectations, and therefore will be acting irrationally according to the situation as depicted in **Figure 2**. Normally B will play the hawk strategy and so will insist on getting the money, so by not paying A will receive less payoff than she would if she complied with the standard equilibrium expectations.

From the perspective of **Figure 2** we can see why A could be rationally criticized for not paying for the beer even if we grant that on this particular occasion she does not have an actual desire to pay for the beer. However, we must emphasize here that there is no *a priori* reason for A to pay for the beer. As noted earlier, there is more than one solution to the problems that are exemplified in **Figures 1** and **2**. The practice that gives rise to the expectations that are captured in Searle's example is a product of the evolution of human practices and societies in general, and in that sense is contingent to the extent that human biological and cultural evolution is contingent. However, if the customer in Searle's example represents a real antisocial personality who does not have any kind of desire or disposition to comply with the social norms that regulate normal behavior in a bar, then we should construe her as having different expectations of how people should behave in these situations, such that the situation will not be properly represented by **Figure 2**. In the latter case, her payoffs should be construed differently because her inclination to play hawk would have to bring her more utility whatever strategy the other

[57] For a defense of a subjectivist account of normative reasons that spells it out in terms of the deliberative point of view of the agent see Arkonovich (2011).

player adopts. In that case, I think we would have grounds for claiming that that person does not really have a (sufficient) reason for complying with the dominant norm (i.e. paying for the beer).

		A	
		Dove	**Hawk**
B	**Dove**	**2** **2**	**3** **1**
	Hawk	**1** **3**	**0** **0**

Figure 2 (adapted from Verbeek, 2007, p. 247)

This depicts the same situation as Figure 1, only the payoffs have been modified to represent the situation in Searle's bar example more closely. If customer A pays for the beer (dove) then bartender B can either take the money (hawk) or, let us say, reduce the price by lowering her margin income (dove). If A refuses to pay, and if B plays hawk, as expected, then neither get anything.

3.6 Concluding remarks and possible objections

In this chapter, I have tried to demonstrate that certain forms of subject-based theories of reasons have the resources to explain why idealization has a natural place inside those accounts. My preferred explanation involves reference to what I have called a version of a response-dependent theory of reasons. I developed my view in analogy with color vision. In this respect I argued that the scientific data coming from studies on certain *types* of people (e.g. VM patients and successful psychopaths) give us reason to believe that the recognition of relations as reasons is partly constituted by the cognitive/affective make up of people and their place in the world.

I will close this chapter with some possible objections to the view of reasons expounded above. This list of objections will serve as an introduction to the topic of the next chapter, where I will try to dispel these objections and provide a broader set of considerations that could serve as a support for the present response-dependentist account of reasons.

Enoch (2005) and other objectivists could complain that the account is too revisionary and that the empirical data I provide in this chapter are not enough to justify the claim that the reasons that matter are those that are in some sense dependent on our cognitive and affective capacities. There are several points on which objectivists might claim that the argument is weak or that I beg the question.

One could follow Parfit (2011a; 2011b) and claim that what matters is an objective thing that reflects the intrinsically valuable, desirable, or reason-providing properties of states of affairs. Furthermore, an objectivist could accept that there are people such as successful psychopaths, who are rational but lack normal affective responses, and still claim that they do not provide an example that counts in favor of a subjectivist view of reasons. The claim could be that our normal affects or emotional reactions are structured in such a way that they track mind-independent normative reality (see e.g. Roeser, 2011). Then the claim would be that successful psychopaths are people who have the same reasons to do things as non-psychopaths, but just do not recognize them.

In effect, the objection could be summarized by saying that the account is revisionary without enough justification. Even in the case of color vision, there are authors who try to save 'common-sensical' mind-independent realism (see e.g. Tye, 2000, chapter 7). So objectivists about normative reasons could similarly claim that there is an *a fortiori* strong reason to save objectivity about normative reasons, especially because their accounts supposedly capture normative phenomenology better than the subjectivist accounts (cf. Enoch, 2005; Parfit, 2011a).

I take these objections as a cue for the topic of the next chapter. In the following chapter, I will introduce broader considerations that should indicate why the previously stated objections do not hold. These considerations should also indicate why some version of the response-dependentist theory of reasons should be favored over object-based theories. The critique will be based on recent naturalistic work that utilizes theories and insights from evolutionary biology and psychology.

4 Normative reasons from an evolutionary perspective

4.1 Introduction

The aim of this chapter is to show that even if we grant that the commonsensical view of normative reasons presupposes mind-independence, the resulting view is not compatible or at least not plausible when evaluated from a scientific point of view that acknowledges the received view from evolutionary approaches to the mind. The position that will be disputed is a robust version of normative realism (FitzPatrick, 2008; Enoch, 2011; Shafer-Landau, 2003). This position can be summarized in three conditions:

1. Normative statements about reasons purport to state facts.
2. At least some normative judgments about reasons are literally true.
3. Truths about normative reasons are stance-independent.

Condition 1) is the familiar idea that normative judgments can be true or false, that is, that they express evaluative beliefs about the world. This view is opposed by non-cognitivists, who contend that normative judgments do not express beliefs but rather some motivational attitude such as desire or states involved in making action-plans (Blackburn, 1998; Gibbard, 1990). Condition 2) states that some of our judgments about normative reality are true. In other words, it states that we have got something right regarding normative reality and that not everything that we believe about normative reasons is false. This condition is rejected by some authors who accept 1). Notably, error-theorists contend that normative judgments purport to state facts but in fact all of them are false when construed literally (Joyce, 2001; Mackie, 1977/1990).

For the purposes of the present chapter, condition 3) plays the most important role because it states that what there is a reason to do is stance/mind or subject independent. To generalize Shafer-Landau, this claim includes the contention that "the [normative] standards that fix the [normative] facts are not made true by virtue of their ratification from within any given actual or hypothetical perspective" (Shafer-Landau, 2003, p. 15). In this chapter I will not directly discuss the plausibility of conditions 1) and 2). Rather I will concentrate on 3) and argue that it cannot be satisfied given evolutionary considerations about the origins and underpinnings of

our judgments about normative reasons. If there are truths about normative reasons they cannot be plausibly construed as completely independent from our actual or hypothetical attitudes.

4.2 Against the mind-independence of reasons: An evolutionary perspective

The evolutionary perspective on normative reasons is most often employed in debunking robustly realist/objectivist positions in metaethics (e.g. Joyce, 2006, chapter 6; Ruse & Wilson, 2006; Street, 2006; 2008b). Moreover, debunking arguments are usually used to undermine a possible *justification* of realist/objectivist claims (see e.g. Brosnan, 2011; Enoch, 2010; Kahane, 2011; Shafer-Landau, 2012). The epistemological construal of evolutionary debunking arguments is well captured in Ruse and Wilson's (2006, p. 566) statement that "even if external ethical premises did not exist, we would go on thinking about right and wrong in the way that we do." We might naturally read this statement as implying that whether moral facts exist or not does not affect the content of our moral beliefs.

Guy Kahane outlines the general structure of evolutionary debunking arguments:

Structure of evolutionary debunking arguments:

1. Causal premiss: Our evolutionary history explains why we have the evaluative beliefs we have.
2. Epistemic premiss: Evolution is not a truth-tracking process with respect to evaluative truth.
3. Metaethical assumption: Objectivism gives the correct account of evaluative concepts and properties.

Therefore:

4. Evaluative skepticism: None of our evaluative beliefs are justified. (Kahane, 2011, p. 115)

The first premiss usually involves giving an evolutionary explanation of the formation or maintenance of evaluative beliefs in the general population of human beings. The second emphasizes the fact that traits evolved because they maximize fitness and not because they reliably track actual states of affairs. The third premiss makes explicit which positions the evolutionary debunking arguments are targeted against. The reason for this is that if we fail to suppose that objectivism or mind-independence are not proper accounts of the evaluative discourse then the argument loses its edge. For example, if we believe that evaluative judgments track truths about our own attitudes or the attitudes we would want ourselves to have when we

are relevantly informed, then the fact that we have evolved to have dispositions to judge in certain ways would not have undermining effects. The reason for this is that the view would be consistent with accepting that what we value depends on our evolved natures.

Finally, the conclusion of the argument states the claim that since evolution is not a truth-tracking process it does not guarantee that the evolved dispositions that influence the formation and maintenance of our evaluative judgments will also track truth about mind-independent reality. Therefore, we cannot be justified in believing that our evaluative judgments, whose formation and maintenance was influenced by evolutionary processes, are epistemically justified.

One instance of this argumentative schema is the following example. Suppose we think that raising one's own children is objectively good, and that therefore everyone has a *pro tanto* reason to take care of his own children.[58] There is a plausible causal-evolutionary story as to why we would have that belief, namely, evolution by natural selection tends to maximize the proportion of those organisms in the population that have greater fitness.[59] In other words, natural selection favors the persistence of those organisms that on average have a greater probability of survival and reproduction, and therefore have greater chances of spreading their genes in the population (Sober, 1999, pp. 58-59). In the case of humans and other mammals, whose survival rates, especially in young age, depend on parents' protection and rearing, the fitness value of their genes will heavily depend on having the disposition to take care of their own children. Therefore, according to this evolutionary explanation, having the disposition to rear one's own children will be beneficial in terms of fitness maximization.

Furthermore, we can suppose that this disposition influenced people with the capacity to form evaluative judgments to offer intuitively compelling judgments of the form: "Taking care of one's own children is good." If the evolutionary explanation of the emergence of the disposition to take care of one's own children is plausible then it also seems plausible that the same disposition can explain the emergence and intuitive appeal of the judgment that rearing one's own children is good. However, now the importance of the second premiss emerges: evolution by natural selection is not a truth-seeking process. What is good for spreading genes in some population or for enhancing the survival and reproductive rates of some organism does

[58] The example comes from Street (2006, p. 115).

[59] The fitness of an individual organism normally refers to the expected number of its offspring that will survive to reproductive age (Garson, 2015, p. 190). Thus, organisms that take care of their offspring will normally increase their own fitness by helping their progeny to reach reproductive age.

not have to reflect true states of affairs in any substantive sense (Stich, 1990, p. 62). On the contrary, believing falsehoods can sometimes be advantageous in terms of fitness maximization. For example, believing that one is professionally extremely competent and very attractive, when this belief is not grounded in facts, could boost one's confidence in such a way that one would on average have more professional and romantic success than a person whose beliefs about herself are grounded in facts.

By combining an explanation of the evolution of the content of some evaluative judgments and the fact that evolutionary processes do not track the truth, we can see why our belief that evaluative judgments represent some objective state of affairs would lose its justification. Such evolutionary explanation also accounts for the fact that we would keep believing that, for example, rearing our own children is good even if there were no objective moral fact ontologically grounding that belief. Thus, the basic idea of epistemologically construed evolutionary debunking arguments is that since the existence or non-existence of moral facts does not affect the actual content of our moral beliefs, we lose the epistemic justification for holding those moral beliefs. From these considerations, some authors conclude that a kind of moral skepticism concerning moral reality is justified (Joyce, 2006). However, a further ontological conclusion, that *there are no moral facts*, would not be warranted because as far as we know moral facts could exist independently of the mind, it is just that we do not *know* whether our moral beliefs correspond to them.

However, evolution-based arguments against objective, mind-independent morality have also been construed as having ontological consequences.[60] This reading of the evolutionary debunking argument is actually endorsed by Ruse and Wilson (2006; see also Rosenberg, 2011, chapter 5):

[60] Joyce (2013) distinguishes between three types of debunking arguments: truth debunking, theory debunking, and justificatory debunking. In the present context truth debunking would refer to the idea that evolutionary considerations show that (all or some subset of) normative claims, even though they pertain to be true, are actually false. Theory debunking aims to show that certain theories about moral judgments are false. This is where the claim that object-based theories of reasons are not compatible or plausible from the perspective of the evolutionary considerations belongs. Justificatory debunking refers to the idea that evolutionary considerations cancel out whatever justification we might have for our normative judgments (or some subset of them). Here is where the already mentioned epistemological construal of the evolutionary debunking arguments belongs. It seems to me that most of the literature concentrates on this third type of argument. However, in this chapter my aim is to consider and defend the second type of (theory) debunking argument that pertains to have ontological consequences, as opposed to narrowly epistemological ones.

> We believe that implicit in the scientific interpretation of moral behavior is a conclusion of central importance to philosophy, namely, that there can be no genuinely objective external ethical premises. (Ruse & Wilson, 2006, p. 565)

I will support this view, because it seems to me that considerations based on the relation between evolutionary theory and normativity have ontological implications for our commonsensical theory of reasons. As far as our commonsense view of normative reasons presupposes or is in some way committed to robust realism about normative facts, I think the commonsense view is wrong. I think that is the case in the specific case of morality, as the above quotation states, and in the more general case of normative reasons.

4.3 Judgments about reasons and their evolutionary underpinnings

In showing why I think object-based theories of reasons are not compatible with a naturalistic world-view, I will heavily rely on an evolutionary argument provided by Sharon Street (2006) against what she calls 'evaluative realism.' There are two reasons for this choice. The first is the fact that Street's argument targets robust realism about normativity in general. So among the targets of the argument are also object-based theories of reasons.[61] The second reason is the apparent contention that Street's argument, if successful, makes it likely that normative facts, if there are any, are grounded in mind-dependent and relational properties of objects (cf. Kahane, 2011, p. 116).

Street (2006) starts her argument with the simple observation that from an evolutionary perspective not all evaluative judgments or attitudes[62] will be on *a par* in terms of fitness benefits. Here is what she says about the possible fitness-detrimental attitudes one could possess:

---

[61] Street (2006, p. 111) follows Shafer-Landau (2003) in claiming that stance-independence is an essential feature of normative realists' views (incidentally Shafer-Landau takes the term from Milo (1995)). Stance-independence refers to the idea that purported normative facts are independent from evaluative attitudes or judgments that a subject could make. I think it is also plausible to identify stance-independence with what Parfit (2011a) calls object-based theories of reasons. According to object-based theories, reasons are grounded in intrinsic features of objects, which means that their existence is independent of any attitude an agent might have towards those objects.

[62] Street (2006, p. 110) subsumes evaluative judgments under the heading of evaluative attitudes. She defines them in the following way: "Evaluative attitudes I understand to include states such as desires, attitudes of approval and disapproval, unreflective evaluative tendencies such as the tendency to experience X as counting in favor of or demanding Y, and consciously or unconsciously held evaluative judgments, such as judgments about what is a reason for what, about what one should or ought to do, about what is good, valuable, or worthwhile, about what is morally right or wrong, and so on."

> It is clear, for instance, how fatal to reproductive success it would be to judge that the fact that something would endanger one's survival is a reason to do it, or that the fact that someone is kin is a reason to harm that individual. A creature who accepted such evaluative judgements would run itself off cliffs, seek out its predators, and assail its offspring, resulting in the speedy elimination of it and its evaluative tendencies from the world. *(Street, 2006, p. 114)*

On the other hand, having opposite evaluative attitudes or tendencies would be beneficial in terms of fitness:

> [I]t is clear how beneficial (in terms of reproductive success) it would be to judge that the fact that something would promote one's survival is a reason in favor of it, or that the fact that something would assist one's offspring is a reason to do it. Different evaluative tendencies, then, can have extremely different effects on a creature's chances of survival and reproduction. *(Street, 2006, p. 114)*

From these observations Street comes to the conclusion that it is plausible that evolution *via* natural selection could have and probably has influenced the contents of evaluative judgments that we currently endorse. In this manner, she proposes to explain why the following judgments about reasons would be endorsed and considered true, at least by members of our species.

> (1) The fact that something would promote one's survival is a reason in favor of it.
>
> (2) The fact that something would promote the interests of a family member is a reason to do it.
>
> (3) We have greater obligations to help our own children than we do to help complete strangers.
>
> (4) The fact that someone has treated one well is a reason to treat that person well in return.
>
> (5) The fact that someone has done one deliberate harm is a reason to shun that person or seek his or her punishment
>
> (6) The fact that someone is altruistic is a reason to admire, praise, and reward him or her. (Ibid. p. 115)[63]

Although Street does not base the acceptance of statements (1)–(6) on any wide-ranging cross-

[63] For a similar set of claims that have intuitive appeal and similar evolutionary explanation see Rosenberg (2011, pp. 65-66).

cultural studies, it is still plausible, given the acceptance of the evolutionary theory, to suppose that people widely endorse these reason-statements. And if that is the case we can ask what explains the emergence of these judgments or attitudes and why we think that there are such reasons? At this point Street (2006) invokes different evolutionary mechanisms that might have influenced the formation of judgments with contents as stated in (1)–(6).

The explanation of (1) seems straightforward. It is plausible that if we care about our own survival then caring about the means that enhance survival will be beneficial for surviving and, at some further future point, for reproducing. Judgments (2) and (3) can be explained by invoking Hamilton's (1964) *kin selection* theory,[64] according to which it can be expected that organisms that share genes will show more altruistic behavior towards each other because such behavior can increase the *inclusive fitness*[65] of those organisms. The classic example is a mother's love for her child. From a genetic point of view, it is clear why parents care for their children, namely by raising and caring for their children they enable their genes to replicate and spread through the population.

Reason-judgments of types (4)–(6) cannot be explained by the theory of kin selection as long as they are taken to apply to non-family members, because benefiting strangers at one's

[64] Hamilton's (1964) theory is most often referred to as *inclusive fitness theory*. The alternative name *kin selection theory* was coined in 1964 by the famous theoretical biologist Maynard-Smith (Rogers, 2010). Recently the significance of the concept of inclusive fitness has been heatedly debated. For example Nowak, Tarnita, and Wilson (2010, p. 1059) claim that inclusive fitness theory is not general enough to provide general dynamics of gene frequencies. In addition, Nowak et al. (ibid.) claim that in limited cases in which the inclusive fitness theory gives right predictions it is equivalent to results derived from standard natural selection theory, and is therefore obsolete. This argument has been disputed by many scientists and theoreticians working in this area of evolutionary biology (for an overview of the debate see Birch & Okasha (2015)). However, the specifics of the debate on the usefulness of the kin selection theory is not essential to the present discussion. For Street's (2006) argument to go through it is enough to show that there is a plausible evolutionary story that can explain why we think certain judgments about reasons are true.

[65] Technically, "inclusive fitness is defined as the sum of the effect of [some] action on the actor's own fitness and on the fitness of the recipient multiplied by the relatedness between actor and recipient, where 'recipient' refers to anyone whose fitness is modified by the action" (Nowak, Tarnita, & Wilson, 2010, p. 1057). Mathematically the notion is expressed as the following inequality: $r > c/b$, where r denotes the relatedness coefficient or, in other words, the probability that the two organisms will share the altruistic gene. c denotes fitness costs to the altruistic actor and b denotes the fitness benefits received by the recipient. The basic idea expressed by the $r > c/b$ inequality is that cooperation will evolve through natural selection if the relatedness between actor and recipient is greater than the ratio between cost and benefit; in other words, if the product of relatedness and fitness benefits is greater than the fitness cost ($r \times b > c$).

own cost will not enhance one's own inclusive fitness value. However, there are five recognized rules by which altruistic behavior could have evolved (Nowak, 2006). First is the above-mentioned kin selection theory. Second is *direct reciprocity*. According to the theory of direct reciprocal altruism, in cooperative interaction one organism – the actor – temporarily incurs fitness costs to itself but increases the fitness benefits of another organism – the recipient – and expects to be repaid from the beneficiary at some later point in time (Trivers, 1971). Since organisms, such as humans, benefit greatly, in terms of fitness, from living in cooperative groups they have an incentive to endorse cooperative or altruistic behaviors and to punish or shun those that are not altruistic (this would be the tit-for-tat strategy (Axelrod, 1984)). For example, I help my neighbor harvest his field and in return expect him to help me harvest my field. If the neighbor does not return the favor, I engage in punitive behavior, such as refusing to help him on the next occasion or in ruining his reputation by spreading news about his non-reciprocal behavior. Therefore, direct reciprocity can plausibly explain the intuitive appeal of judgments (4) and (5).

What about situations in which we do not expect to be reciprocated, because, for example, there is little chance of encountering the person again? For example, we donate money to charity without the expectation of being repaid by the people that receive our charity. Similarly, we often feel that a person should be punished in some way even though she has not done us any direct harm. Moreover, we often go out of our way to punish her even though the costs of punishment outweigh the direct potential benefits.[66] In other words, people are apparently prone to strongly reciprocal behavior.[67] To explain this kind of altruistic behavior the mechanism of indirect reciprocity is invoked. Here the most important concept is that of reputation (Sperber & Baumard, 2012):

[66] There is empirical evidence supporting the claim that people exhibit strong altruistic (reciprocal [see footnote 67]) behavior even when there is no foreseeable possibility of being reciprocated (Fehr & Fischbacher, 2003). Moreover, other studies provide evidence that people often engage in punishing behavior that is also costly to the punisher with no expectation of being repaid (Fehr & Gächter, 2000; 2002).

[67] Strong reciprocity is defined as "a combination of altruistic rewarding, which is a predisposition to reward others for cooperative, norm-abiding behaviours, and altruistic punishment, which is a propensity to impose sanctions on others for norm violations. Strong reciprocators bear the cost of rewarding or punishing even if they gain no individual economic benefit whatsoever from their acts" (Fehr & Fischbacher, 2003, p. 785). Strong reciprocity is contrasted with (weak) reciprocal altruism, according to which altruists "reward and punish only if this is in their long-term self-interest" (ibid.).

> Helping someone establishes a good reputation, which will be rewarded by others. When deciding how to act, we take into account the possible consequences for our reputation. We feel strongly about events that affect us directly, but we also take a keen interest in the affairs of others, as demonstrated by the contents of gossip. *(Nowak, 2006, p. 1561)*

By being helpful across various situations and towards different people one can build up one's reputation in a way that can compensate for the many costs incurred by this altruistic behavior. For example, studies have shown that people who are more helpful get a positive reputation and in effect receive more benefits in return (Wedekind & Milinski, 2000). Indirect reciprocity can explain the intuitiveness of judgments (4), (5), and (6) when they are construed as judgments about people with whom we do not necessarily interact and reciprocate directly.

The fourth rule is so-called *network reciprocity* (Nowak, 2006, p. 1561; see also Skyrms, 1996). It builds on the aforementioned accounts by noting that altruistic behavior will evolve and be maintained as a result of the structure of the population. Networks represent the neighborhoods within which an agent acts. In these models, agents "observe the action and resulting payoff of their neighbours and preferentially imitate the action played by high-payoff individuals" (André & Morin, 2011, p. 2532). Altruism will emerge if altruistic agents receive on average more fitness benefits than non-altruists. This network-based selection has the potential to explain the ubiquity of judgments (2)–(6) if holding or being disposed to hold these judgments on average provides the highest amount of fitness benefits to actors in the network.

The fifth rule relies on the idea of group selection. According to this concept, the evolution of altruism can be explained by natural selection working at the level of a group of organisms. Unlike direct and indirect reciprocity, which are paradigmatic examples of selection at the individual level, group selection relies on the idea that selection happens on multiple levels; it depends on selection between individual organisms and between groups of organisms (Okasha, 2006; Sober & Wilson, 1998). Here the basic idea is that altruists will tend to form groups that, because of their greater fitness, grow and split faster than other groups that contain non-cooperators (i.e. free-riders).[68] Since nature's resources are finite, the group with a faster

[68] The basic mechanism that explains why altruistic groups will split faster and therefore tend to predominate is the following. According to group selection theory, groups are dynamic entities, which means that they dissolve into greater populations and then recongregate again into smaller groups. In this process, the main supposition is that altruists will tend to form clusters in which they cooperate and share the benefits of cooperation. Since in groups of cooperators fitness benefits are greater than in groups that have many non-cooperators, the former groups' offspring will increase in number, while the number of non-cooperators will tend to decrease. In effect, the growing number of cooperators will make it the case that in the next splitting and reforming of the smaller

splitting rate will tend to predominate and force other competitor groups towards extinction (Nowak, 2006, p. 1561). Group selection plausibly explains the intuitive appeal of judgments (2)–(6). However, it seems especially potent to explain the intuitiveness of moral judgments of type (6). The reason for this intuitiveness seems to be the intrinsic value that people normally associate with living in functional communities (Haidt, 2007). The support and the encouragement of altruistic behavior captured in (6) is certainly well explained by a psychologically favorable disposition towards conspecifics or in-group members that is underpinned by the competition (selection) between groups (Garson, 2015, p. 34).

In this section, following (Nowak, 2006), I have introduced five mechanisms for the evolution of altruism that can explain the intuitive appeal of different widely endorsed judgments about reasons (such as those captured in statements (1)–(6)). I enumerated these mechanisms as if they formed complementary explanations. However, in certain forms these mechanisms are often used to provide competing explanations of the evolution of cooperation. So, by extension they could also be used as competing explanations of the intuitive plausibility of judgments (1)–(6).[69] However, in this chapter I do not aim to adjudicate between the plausibility of evolutionary explanations that invoke different mechanisms. For our present purposes, it is enough to recognize that the evolutionary theory has the resources to explain how the contents of our evaluative judgments came about (Krebs, 2011).[70] That is, it has the

groups a greater number of cooperators will cluster into new groups of cooperators, and by iteration this process could at the limit lead to the extinction of non-cooperating individuals (Garson, 2015, pp. 34-35).

[69] For example, proponents of the idea that people are strongly altruistic (reciprocal) tend to think that cooperation between humans is best explained by a cultural version of group selection theory (Boyd & Richerson, 1985; Fehr & Fischbacher, 2003; Gintis, 2003). Cultural group selection refers to the group-beneficial selection of traits (such as social behaviors) that are culturally transmitted through different learning mechanisms (most notably different forms of imitation). On the other hand, other authors in the area are more critical of the importance of cultural group selection (for a critical discussion see André & Morin (2011) and Morin (2014)) and argue that forms of indirect reciprocity are sufficient to explain complex forms of altruism (see e.g. Baumard, André, & Sperber, 2013).

[70] It is important to emphasize that Street does not claim that the concrete contents of our judgments about what we have or do not have a reason to do are somehow coded in our genes and then passed from one generation to the next. What is deeply rooted in our evolutionary history is what Street (2006, p. 119) calls evaluative tendencies, that is, an "unreflective, non-linguistic, motivational tendency to experience something as 'called for' or 'demanded' in itself." These evaluative tendencies we share, to some degree, with our primate relatives (see de Waal, 1996). What is important here is Street's claim that evolutionary processes *directly* influence the shape of these basic evaluative tendencies "and that these basic evaluative tendencies, in their turn, have had a major

resources to explain how judgments about what we think we have a reason to do and care about have been influenced by the workings of natural selection.

4.4 Street's Darwinian dilemma for a normative realist

Now we have enough evolutionary background to formulate Street's (2006) dilemma for an evaluative or normative realist.[71] It seems undeniable that evolutionary forces had some impact on the contents of our normative judgments. The question is: what is the relation between the fact that evolution influenced the formation of many normative judgments we accept and the posited structure of independent normative truths? The normative realist has two options: either she can claim that evolutionary processes that influenced which contents we affirm in our normative judgments *do not stand in any relation* to independent normative truths or she can claim that they *stand in some kind of relation* to independent normative truths.

Street swiftly discards the first horn of the dilemma as implausible.[72] The reason seems to be the following consequence: if someone wants to claim that evolutionary processes have not favored the emergence of the capacity to grasp the true contents of normative reasons, then the evolutionary forces that actually have influenced which normative judgments we accept must be viewed as a purely distorting factor with respect to the tracking of the normative truth (cf. Street, 2006, p. 121). Here is how Street argues against the first horn of the dilemma:

> On this view, allowing our evaluative judgements to be shaped by evolutionary influences is analogous to setting out for Bermuda and letting the course of your boat be determined by the wind and tides: just as the push of the wind and tides on your boat has nothing to do with where you want to go, so the historical push of natural selection on the content of our evaluative judgements has nothing to do with evaluative truth. Of course every now and then, the wind and tides might happen to deposit someone's boat on the shores of Bermuda. Similarly, every now and then, Darwinian pressures might have happened to push us toward accepting an evaluative judgement that accords with one of the realist's independent evaluative truths. But this would be purely a matter of chance, since by hypothesis there is

influence on the evaluative judgements we affirm" (Street, 2006, p. 120). So the claim is that evolutionary processes *indirectly* influenced the formation of the actual contents of our normative judgments.

[71] In what follows, 'normative realist' refers to someone who accepts robust normative realism as defined in the first section of this chapter.

[72] I write 'swiftly' because, compared to the second horn, Street gives relatively little space to the discussion of the first horn of the dilemma (only 3 of 58 pages; see Street (2006, pp. 121-125)).

> no relation between the forces at work and the 'destination' in question, namely evaluative truth. *(Street, 2006, pp. 121-122)*

Against the plausibility of the first horn of the dilemma, Street explicitly utilizes the epistemological consideration that corresponds to the second premiss of the general schema of evolutionary debunking argument (Kahane, 2011). The claim is that to the extent that evolution by natural selection has shaped the content of our evaluative judgments, it has had a distorting influence concerning the reliability of our evaluative judgments, because evolutionary processes generally do not track the truth (Stich, 1990). Thus, Street claims that it would be a massive coincidence if the evaluative beliefs we have ended up with as products of blind evolutionary processes were exactly those that reflect the mind-independent structure of normative reality. As we have seen, evolutionary considerations can explain why we think that evaluative judgments with particular contents are true, and because of this we need to be suspicious about their reliability. If we adopt the first horn of the dilemma, the belief that we have a reason to care about our children, for example, would be undermined, because there is a clear evolutionary explanation for holding a belief with the latter content. And this seems to be a very unintuitive result. According to this line of reasoning, "we are left with the implausible skeptical conclusion that our evaluative judgements are in all likelihood mostly off track, for our system of evaluative judgements is revealed to be utterly saturated and contaminated with illegitimate influence" (Street, 2006, p. 122).

A proponent of the first horn of the dilemma may object to Street's argument by claiming that even though evolutionary processes influenced the contents of our normative beliefs, people have other evolved mental faculties through which they can reach independent normative truths (see e.g. de Lazari-Radek & Singer, 2012, p. 16). The usual claim is that people have the capacity to reason, which enables us to reflect on our evolutionary, culturally, experientially, etc. given evaluative beliefs. The capacity to reason enables people to abandon, revise, or confirm their non-rationally acquired beliefs and therefore the possession of this capacity explains and justifies the process of acquiring, abandoning, or revising evaluative beliefs.

Street (2006, pp. 121-125) anticipates this objection and answers it by acknowledging that we cannot give a full explanation of how people acquire their evaluative beliefs if we do not recognize that reasoning and reflection on one's beliefs and attitudes can influence one's evaluative beliefs. However, Street points out that the problem of explaining the massive coincidence between people's evaluative beliefs and the contents of normative reality is not solved just by introducing the ability to reason and reflect on our attitudes. Street (ibid., p. 124)

points out that rational deliberation or reflection starts by presupposing some evaluative judgments; the process does not happen in a deliberative vacuum. When we rationally try to reach some decision or decide whether some evaluative belief is justified we hold fixed some evaluative and factual judgments in the background and from that point of view try to evaluate the original beliefs. Since we cannot survey and reflect on every normative judgment that we hold, it is likely that those judgments that form the background of our deliberations are those that are deeply entrenched in our psyche. It is plausible to suppose that exactly those entrenched judgments (or the underpinning evaluative dispositions) emerged as a consequence of Darwinian selection processes.

One could claim that we learn about and respond to normative facts by *rationally perceiving* those facts as reason-giving. However, this model of rational perception cannot be accommodated in terms of ordinary, empirical perception, because by the original construal of normative realism, reasons are further properties of non-normative facts. Thus, grounds of reasons can be empirical facts, such as the fact that the streets are wet; however the fact that the streets are wet, according to this view, has (or could have) the further property of providing a reason to believe that rain has been falling. Therefore, a model of rational perception should include some foundational normative intuitions that ground other evaluative beliefs (see e.g. Huemer, 2005). Therefore, according to the evolutionary argument, these foundational intuitions would be the likely products of evolutionary forces.

Thus far the argument has been that if we grant that normative intuitions are influenced and formed on the basis of evolutionary processes and that those processes do not have any connection to normative reality, then it would be a massive coincidence if our 'rational intuitions' really corresponded to some basic normative facts. To remedy this problem the natural move is to claim that rational reflection involves an attempt to reach reflective equilibrium between our normative intuitions, factual beliefs, and more general evaluative beliefs (see e.g. Daniels, 1996). The idea is that through rational reflection, by confronting evaluative beliefs against factual and other normative beliefs, we are purged of unjustified beliefs (the remnants of our naturally acquired predispositions), leaving only justified beliefs about normative reality. However, as Street (2006, p. 124) points out, the fact that we could reach reflective equilibrium concerning our evaluative beliefs still does not answer the coincidence problem. If our intuitions and evaluative beliefs are largely products of evolutionary processes, then it is not clear why the final product would be any more likely to reflect the real contents of normative reality as opposed to biases that reflect fitness considerations. Since reflective equilibrium is a coherentist procedure, "we can test our

evaluative judgements only by testing their consistency with our other evaluative judgements, combined of course with judgements about the (non-evaluative) facts" (ibid.). Therefore, Street concludes that:

> if the fund of evaluative judgements with which human reflection began was thoroughly contaminated with illegitimate influence – and the objector has offered no reason to doubt this part of the argument – then the tools of rational reflection were equally contaminated, for the latter are always just a subset of the former. *(Street, 2006, p. 124)*

Let us observe the argument so far. Street claims that by endorsing the first horn, normative skepticism will be warranted, given the fact that evolution influenced the formation of our normative beliefs. Most opponents of Street's argument take issue with the claim that the introduction of evolutionary considerations turns the correspondence between normative judgments and mind-independent normative reality into a massive coincidence (Brosnan, 2011; Enoch, 2010; Kahane, 2011; Parfit, 2011b; Shafer-Landau, 2012; Skarsaune, 2011; de Lazari-Radek & Singer, 2012). This in effect amounts to accepting the first horn of the dilemma. I do not have space to review all of the written responses to Street's contention that acceptance of the first horn leads to an overarching skepticism regarding normative knowledge (robustly construed). Certainly, among these responses there are good explanations of how robust normative realism might accommodate the existence of normative facts and the workings of natural selection.

However, I think the sharpness of Street's (2006) argument lies somewhere else.[73] The construction of the argument in terms of a dilemma allows us to distinguish those who are naturalistically inclined, in the sense of accepting the authority of the current body of scientific knowledge and explanations, and those who are not so inclined. Endorsement of the first horn delineates those who are not naturalistically inclined. The reason for this is the following: current evolutionary theory has the resources to explain the emergence of the contents of widely endorsed normative judgments and further, to explain why we take those contents to be intuitive and true. But the acceptance of the first horn in effect denies the plausibility of those explanations by claiming that there is no explanatory relation between the evolution of normative tendencies and the reason why we take the judgments that stem from these evolved tendencies to be true.

[73] I do not want to claim that what I say in the main text is something that Street would accept as a valid interpretation or the aim of her argument.

Therefore, someone who accepts the first horn in effect denies the plausibility of evolutionary explanations of the emergence of normative judgments and in that respect she might be deemed scientifically skeptical.[74] However, my reading of Street's argument is that it is targeted against those who are naturalistically inclined and not scientific skeptics. If someone does not accept evolutionary theory or the way it explains the emergence of normative phenomena then the evolutionary debunking argument will have no effect on her.[75] Since naturalistically inclined authors will negate the first horn of the dilemma this brings our discussion to the second horn of the dilemma.

However, before I address issues concerning the second horn of the dilemma I want to discuss one possible objection that in one sense relates to the first horn of the dilemma and in another dismisses the dilemma as unfounded. The objection is that Street's argument rests on a mistake. The mistake is the belief that normative judgments about what we have a reason to do or believe have or had adaptive value.[76] So the driving force behind the objection is that evolutionary considerations do not explain why we find certain basic normative judgments plausible and intuitive.

---

[74] Someone might argue that I put an unjustified burden on the robust realist who endorses the first horn to accept methodological (and the ensuing ontological) naturalism when there is in fact a legitimate philosophical position according to which philosophy has a certain amount of autonomy to pursue its own domain of inquiry that cannot be encompassed with empirical methods (see e.g. Shafer-Landau, 2012; Smith, 1994). For example, it could be argued that the ultimate metaphysical reality cannot be investigated with empirical methods, because every metaphysical theory will be underdetermined by the empirical evidence, theories, and methods that we possess. My reply to this objection is that even if we admit that this view is plausible about general metaphysics, it does not apply to the present case. The reason for this is that with respect to normative reality we have good and plausible explanations that are or could be derived from accepted scientific theories. So there is no pressing need to adopt a non-naturalistic methodology and delineate a specifically philosophical domain of inquiry from all other scientific domains. Therefore, in this context I think that those who reject the possibility of a scientific explanation of the evolution and emergence of normative phenomena in the practical domain could legitimately be deemed scientific skeptics.

[75] In this respect I follow Rosenberg's (2011) line of thought. In one of his talks that I attended in Prague in 2014, Rosenberg explicitly stated that his evolutionary debunking argument was targeted against those normative realists who are also naturalists – presumably because one cannot expect other non-naturalistically inclined authors to accept the presuppositions of a naturalistically based argument.

[76] The following objection is based on considerations that Parfit (2011b, sections 117-118) uses to argue against Street's evolutionary argument.

4.4.1 From motivations to evaluative beliefs

According to object-based theories of reasons, facts about reasons are facts about the normative properties of non-normative facts, state of affairs, objects, etc. For example, the fact that A is in pain is a descriptive fact about A. On the other hand, the fact that A's being in pain gives her a reason to avoid the source of her pain is a normative fact. In Parfit's (see e.g. Parfit, 2011b, pp. 505, 529-530) terms, the property of being in pain provides A with another, further property of having a reason to avoid the source of the pain. Street's debunking argument seems to presuppose that the belief that being in pain provides one with a reason to avoid pain had some adaptive value among our ancestors. At this point one might object that what actually had adaptive value, that is, what promoted survival, was the *motivation* to avoid pain and not a further belief that pain provides *a reason* to avoid painful stimuli. And since only the motivation to avoid pain is directly advantageous the fact that having this motivation led early humans to believe or judge that being in pain is bad cannot be explained in terms of fitness values (Deem, 2016).

According to Parfit, since the belief that pain is bad "was not advantageous, we have less reason to assume that we would have formed this belief whether or not it was true" (2011b, p. 529). Therefore, the skeptical conclusion is undermined, because the argument's premise – according to which the content of evaluative beliefs can be given an evolutionary explanation – is unsound.

However, this objection is not very persuasive. Parfit also admits that as far as we know the adaptive disposition to avoid painful stimuli "*led* later humans to believe that we have this reason" to avoid painful stimuli (ibid., emphasis added). If adaptive motivations caused humans to adopt evaluative beliefs that reflect those motivations then surely we would have formed those evaluative beliefs whether or not they were true. After all, if evolutionary considerations establish reasonable doubt about the capacity of our evolved motivations to track independent normative reality then whatever they cause will also be susceptible to the same skeptical doubts. Street's (2006) argument only presupposes that evolution had some *indirect* impact on the content of our evaluative judgments (see footnote 70), and this is enough to launch the epistemological worries.

One could try to deny that there is a connection between our evolved motivational dispositions and rationally acquired evaluative judgments. The idea might be that rational people somehow manage to get rid of the evolutionary baggage through the use of reason. However, even if such a thing is possible, just stating the possibility is not enough to answer

the skeptic's doubts. First, it would not answer the question of how people's intuitively plausible evaluative judgments come to be aligned so well with motivations that have evolutionary-based explanations. Here the normative realist would again be faced with the Darwinian dilemma and forced to respond to it.

Second, and more importantly, there is strong empirical evidence from social and moral psychology that supports the view according to which emotional, intuitive, and often unconscious motivations cause or in some other way influence the formation of our more conscious and reflective judgments (see e.g. Haidt, 2001; Nisbett & Ross, 1980). For example, people's first impressions influence their evaluations: "positive evaluations of non-moral traits such as attractiveness lead to beliefs that a person possesses corresponding moral traits such as kindness and good character" (Haidt, 2001, p. 82). Another illustrative example that involves the primacy of the intuitive is 'moral dumbfounding' (Haidt & Bjorklund, 2008; Hauser, Young, & Cushman, 2008). When asked to judge whether consensual safe sex between siblings is wrong, most people, based on their aversion to this act, judge the action to be wrong even though their reasons for thinking so do not track any supposed features of the imagined situation (such as the possibility of getting pregnant, getting some disease, etc.). In the most extreme case it has been shown that when people are primed to feel disgust when reading a text that includes an otherwise neutral word such as 'often' or 'take' they are more likely to provide harsher moral judgments of the acts of the person with whom the primed words are associated (see Wheatley & Haidt, 2005). In general, when people's *intuitive* worldviews are contested by providing divergent evidence people normally engage in reasoning that reflects their intuitive theoretical or practical commitments (cf. Haidt, 2001, p. 821). This primacy of intuitive, automatic, and affective dispositions makes evolutionary sense, because it is plausible that they are the first to emerge in phylogenetic and ontogenetic development and aid adaptation to natural and social environments (Haidt, 2001, p. 819; see also Krebs, 2011).

Therefore, there is persuasive empirical evidence showing that reflective judgments and reasoning are more likely to reflect our evolved motivational dispositions and provide rationalizations of them (cf. Haidt, 2001) than some objective, mind-independent reality (see also Braddock, 2016). Street's claim that human rational and reflective capacities will be as contaminated by evolutionary forces as our more intuitive and motivationally based dispositions seems to be vindicated by broader scientific considerations based on empirical evidence. So if we want to avoid the claim that evaluative human judgments hit upon normative truth just by sheer accident, given the fact that our motivational and cognitive capacities are

products of evolutionary processes, it seems that the only available move for a normative realist is to accept the second horn of the dilemma.

4.4.2 Normative and descriptive fit: Tracking or emergence of reasons?

Accepting the second horn of the dilemma means claiming that there is some kind of positive relation between evolutionary forces and mind-independent normative truth. Street (2006, p. 121) writes that the natural thing for a normative realist to do is to construe the positive relation between evaluative facts and evolutionary forces as a *tracking* relation. The normative realist can say that evolutionary causes track normative truths because, at some point in human history, it was advantageous (in terms of survival and reproduction) to be motivated to act in accordance with mind-independent normative truths. Street calls this proposal the *tracking account* (ibid., p. 126).

To illustrate the tracking account Street reports Parfit's view on the positive relation between evaluative truths and evolutionary forces. However, as we will see, Parfit's published views introduce a little more complexity into the present discussion than is envisaged by Street's (2006) argument.[77] In his book, Parfit says the following about the relation between evolutionary forces and evaluative facts:

> Natural selection slowly but steadily gave later humans greater cognitive abilities. Just as the faster cheetahs and taller giraffes tended to survive longer and have more offspring, who inherited similar qualities, so did the humans who were better at reasoning validly and responding to reasons. *(Parfit, 2011b, p. 494)*

Street interpret the quote such that it provides us with the tracking account, claiming that Parfit provides the hypothesis that "our ability to recognize evaluative truths, […] conferred upon us certain [fitness] advantages" and that "the evaluative judgements that it proved most selectively advantageous to make are, in general, precisely those evaluative judgements which are true" (Street, 2006, p. 126). So the idea is that there is a conjunction of two claims: "the widespread presence of some evaluative judgments rather than others in human population […] is explained

[77] I do not mean this as a criticism of Street's (2006) paper because the view that she quotes from Parfit is taken from their personal correspondence (see ibid., endnote 27), and Parfit's published views on the topic came 5 years later (see his 2011b, chapter 33). However, I believe that Street (2006) has already given all the material needed to answer criticisms or possible responses to her argument from Parfit.

by the fact that these judgments are true" and "that the capacity to discern such truths proved advantageous for the purposes of survival and reproduction" (ibid.).[78]

Before proceeding, it is worth recapitulating what is at stake here. It is plausible to say that the basic premiss of evolutionary approaches to the mind is that people ultimately act *as if* they aim at maximizing their inclusive fitness (El Mouden, Burton-Chellew, Gardner, & West, 2012; Garson, 2015).[79] If the ability to respond to reasons was selected for, this must have been because it played some role in maximizing inclusive fitness among those who developed such a capacity. However on a robust realist's interpretation, 'responding to reasons' cannot be construed as responding to whatever played a beneficial role in the evolution of fairness, altruism, cooperation, and morality more generally. According to a proponent of robust realism, 'responding to reasons' refers to responding to a preexisting normative reality, and therefore she must claim that this capacity was selected for because responding to *specific* independently existing normative truths increased the inclusive fitness of those with that specific capacity. This idea is captured by the tracking account. In Street's words:

> [I]f it is asked why we observe widespread tendencies to take our own survival and that of our offspring to be valuable, or why we tend to judge that we have special obligations to our children, the tracking account answers that these judgements are true, and that it promoted reproductive success to be able to grasp such truths. *(Street, 2006, p. 126)*

However, when normative realists accept the scientific explanation of the link between the content of our normative judgments and the evolutionary processes they expose themselves "to

[78] Parfit could be interpreted as endorsing only the second conjunct. For example, he says that "normative beliefs would have seldom been advantageous" (Parfit, 2011b, p. 514, see also pp. 527-528). If he does not think that normative beliefs were the target of selection then presumably he does not think that their truthfulness plays any significant role in explaining their selective advantage either. The reason for this is that for the ability *to respond to reasons* to evolve, "early humans did not need to have the concept of a normative reason, nor did they need to have normative beliefs about such reasons" (ibid., p. 515). However, I will follow Street (2006) in construing the normative realist as accepting both conjuncts, because it is not clear what explanation could be given for the evolution of the capacity to respond to robust normative reasons if the presupposition of their truth or existence does not play a role in the explanation of the evolution of the capacity.

[79] This does not mean that people consciously act so as to maximize their inclusive fitness. Here the distinction between proximal and distal mechanisms is important. In our case proximal mechanisms are constituted by whatever psychological mechanisms motivate people's behavior. For example, it is widely agreed that people act out of altruistic motivations. Distal mechanisms refer to explanations of how those proximal mechanisms arose and spread among people. For example, kin selection theory explains why people act altruistically in a wide range of situations.

all the usual standards of scientific evaluation, putting [their account] in direct competition with all other scientific hypotheses as to why human beings tend to make some evaluative judgements rather than others" (Street, 2006, p. 126).

In this context, Street opposes the tracking account to what she calls the *adaptive link account* (ibid., p. 127). According to this account, the tendency to produce evaluative judgments with certain kinds of contents evolved "not because they constituted perceptions of independent evaluative truths, but rather because they forged adaptive links between our ancestors' circumstances and their responses to those circumstances, getting them to act, feel, and believe in ways that turned out to be reproductively advantageous" (ibid.). By 'adaptive links' Street means the mechanisms that connect certain circumstances in which an organism evolved and the (behavioral) responses that promoted or still promote the fitness benefits of that organism. In this respect the adaptive link account is an instance of an explanation from evolutionary psychology that utilizes the idea of a modular mechanism.[80]

As examples of such adaptive mechanisms we can mention two cases. The feeling of disgust is taken to have an evolutionary adaptive explanation. The widely held explanation is that disgust is an adaptive response to "substances that might cause illness" (Curtis, Aunger, & Rabie, 2004, p. S131), such as rotten meat, faeces, vomit, maggots, lice, worms, rats, bodily fluids, etc. So the mechanism of disgust forms an adaptive link between circumstances that involve one of the above-mentioned substances and a response that outputs facial expressions and activates behavioral patterns to avoid the substances causing the disgust reaction. The other example provided by Street (2006, p. 127) is an adaptation to avoid painful stimuli. When a person touches the hot stove, a reflex is activated to withdraw the hand from the stove. Here the automatic mechanism is adaptive because it connects important – fitness diminishing – circumstances with adaptive responses, that is, with responses that avoid fitness-diminishing stimuli.

Analogously, Street claims that evaluative judgments could be compared to the latter kind of mechanisms. On the *adaptive link* account, evaluative judgments, and more primitive dispositions that give rise to them, are important because they connect evolutionary significant circumstances in which an organism finds itself and adaptive responses. As mentioned, a judgment that there is a reason to take care of one's own children reflects a fitness-enhancing

[80] In evolutionary psychology, modules are domain-specific mechanisms that evolved as solutions to a particular problem that was recurrent in the environment of evolutionary adaptation. Domain-specificity refers to modules' being sensitive to a particular sort of input from the environment.

behavioral strategy by reinforcing the motivations that one already has when it comes to rearing one's own children. Similarly, judging that there is a reason to reciprocate when someone does you a favor reflects a reinforcement of motivations to endorse reciprocating behavior because it had or has fitness-beneficial consequences. The adaptive link account also has the capacity to explain emotions and their role in normative judgments. For example, it is plausible that the strong feelings that parents have for their children evolved in order to directly protect their inclusive fitness. Furthermore, moral emotions, such as resentment and righteous anger, evolved in order to protect individuals from possible cheaters who take advantage of cooperation but do not reciprocate (Gaus, 2011). In addition, emotions such as the feeling of guilt evolved to regulate behavior in order to overcome incentives to gratify desires for short-term benefit as opposed to long-term benefits. All these emotions underpin normative judgments that we use to criticize others' and our own behavior in order to reinforce the behaviors that have adaptive value.

According to Street, judgments about reasons can to a large degree be construed as adaptive mechanisms albeit more flexible than reflexes (such as disgusting reactions and body reflexes): "From an evolutionary point of view, each may be seen as having the same practical point: to get the organism to respond to its circumstances in a way that is adaptive" (Street, 2006, p. 128). However, unlike hard-wired reflexes, normative judgments can be construed as more conscious and plastic responses to facts that are experienced as *calling for*, *demanding*, or *counting in favor of* certain kind of response (ibid.).

To illustrate the difference between the tracking and the adaptive link accounts Street offers the following examples:

> Consider, for instance, the judgement that the fact that something would promote one's survival is a reason to do it, the judgement that the fact that someone is kin is a reason to accord him or her special treatment, and the judgement that the fact that someone has harmed one is a reason to shun that person or retaliate. […] According to the tracking account, however, making such evaluative judgements contributed to reproductive success because they are true, and it proved advantageous to grasp evaluative truths. According to the adaptive link account, on the other hand, making such judgements contributed to reproductive success not because they were true or false, but rather because they got our ancestors to respond to their circumstances with behavior that itself promoted reproductive success in fairly obvious ways […]. *(Street, 2006, pp. 128-129)*

Now that we have the two competing accounts of how 'responding to reasons' evolved on the table, we can examine their scientific merits. It is to this task that I turn in the next subsection.

4.4.3 The tracking account and the adaptive link account: Scientific merits

As scientific accounts, the tracking and the adaptive link explanations are susceptible to the usual criteria of scientific evaluation. According to Street, there are three relevant criteria that can be used to adjudicate between them. These include parsimony, explanatory clearness, and explanatory unification. According to these criteria, the adaptive link account fares better (cf. ibid., p. 129). Let us compare the cases by utilizing the proposed criteria.

(1) The adaptive link account is more parsimonious because, unlike the tracking account, it does not posit the existence of independent normative truths that our evolved normative judgments track. Saying that the endorsement of specific normative judgments was advantageous is saying that those judgments had favorable effects on an organism's survival and reproductive chances; adding that those judgments had these effects *because* they are *true* does not improve the explanation of why a certain trait evolved as it did, it just makes the hypothesis non-parsimonious.

(2) The tracking account is less explanatorily clear because it does not explain how the *truthfulness* of certain evaluative judgments could explain their reproductive success. Since the content of evaluative judgments is fairly abstract, unlike perceptual content about predators, food resources, shelters, etc., it is obvious that a creature "can't run into such truths or fall over them or be eaten by them" (ibid., p. 130). It is therefore not clear why or how the existence of mind-independent evaluative facts could explain the fitness benefits of the organisms that are able to perceive them (see also Parfit, 2011b, p. 514 for a similar opinion).

(3) The adaptive link account provides unification of disparate phenomena that cannot be accounted for in terms of the tracking account. For example, it has a unified explanation according to which judgments about reasons regarding self-interest, family members, non-relatives, the relative importance of people to other forms of life, etc. fall under a single principle: "they forge links between circumstance and response that would have been likely to promote reproductive success in the environments of our ancestors" (Street, 2006, p. 134). The tracking account can only say that all unrelated kinds of judgments about reasons are adaptive because they are true. The latter statement is as explanatorily non-illuminating as saying that God created every species separately and adapted it to a particular environment, despite the fact that we have overwhelming evidence that all species share common biological roots.[81]

Furthermore, the adaptive link account explains why out of the many logically possible contents that our normative judgments could have acquired, only some particular subset of those

[81] This is the so-called tree of life hypothesis (see e.g. Sober, 1999).

judgments is actually endorsed by human beings. For example, according to the adaptive link account it is clear why we think that infanticide is not commendable, why plants are not more valuable than people, or why the fact that arsenic is poisonous is a reason not to digest it: "namely, […] such judgements – or evaluative tendencies in these general sorts of directions – forge links between circumstance and response that would have been useless or quite maladaptive as judged in terms of reproductive success." On the other hand, "[t]he tracking account has nothing comparably informative to say. It can just stand by and insist that such judgements are false – reaffirming our convictions but adding nothing to our understanding of why we have them" (ibid.).

Additionally, I will argue that the tracking account is less plausible than the adaptive link account because (a) it misconstrues the role of evaluative beliefs in an organism's behavioral and mental economy, and (b) it wrongly portrays the nature of evolution by natural selection as a process that strives towards some ultimate, pre-specified goal, namely towards tracking what ought to be done. I will start with the latter point.

Saying that it was advantageous to form an evaluative judgment with content X because X is true is tantamount to saying that judgment that X was advantageous to form because it ought to be the case that what X claims ought to be the case. This latter claim seems to have strange anti-naturalistic consequences: it imports strange teleological considerations into an evolutionary account. To illustrate the point let us suppose that X stands for *the fact that p is a reason to Φ*. Claiming that it was advantageous to judge that p is a reason to Φ is tantamount to claiming that the latter kind of judgment evolved because it was (is) *true* that p is a reason to Φ.[82] Using a simple rule of truth predicate elimination we can infer the following statement:[83] it was advantageous to judge that p is a reason to Φ because p *is* a reason to Φ.

[82] Here one might also argue against the tracking account by saying that even if natural selection tracked normative reality, this does not mean that what was normatively true in the evolutionary adaptive environment is is still true today. That is, even if the judgment that p is a reason to Φ was advantageous in the past because it was true, this does not mean that it would be advantageous in the present because it is true or that it is even true that p is a reason to Φ (see Street, 2006, p. 133). However, I will not pursue this line of argument further, because I believe there are other more serious objections against the realist's ontology of normative facts (see the main text above).

[83] The truth predicate has two basic inferential rules that govern the application of the term. The introduction rule says that if S is the case then it is true that S [S⊢T(S)]. The elimination rules says that if S is true then S is the case [T(S)⊢S]. For the purposes of my argument the elimination rule is important. In this instance it has the following content: '*it is true* that fact p is a reason to Φ' ⊢ 'fact p is a reason to Φ.'

More generally, according to this proposal, we can say that judging that there is a reason to Φ was advantageous because there is a reason to Φ. Since reasons are normally taken to have certain direction and recommending force the latter statement implies that natural selection, in this case, is directed by external facts about what there is a reason to do or what ought to be the case. And this seems to be an odd result with respect to the fact that natural selection is supposed to mechanically act on blindly generated variation among traits.[84] This view implies the existence of an alignment between the content of a normative judgment and mind-independent normative facts, either because evolutionary processes reached alignment by *trial and error* or because evolutionary processes literally *tracked* normative facts.

The first disjunct has already been discredited – in the discussion of the first horn of Street's (2006) dilemma. The second disjunct is implausible precisely because in the case of normative facts it implies that evolution by natural selection is a goal-directed or purposeful mechanism. This conclusion follows from the following considerations: according to object-based theories, the supposed normative truths to which normative judgments refer are non-causal (Parfit, 2011a; 2011b). Natural selection filters out traits that are less fitness beneficial in comparison to the average fitness levels of those traits in the population. In order for traits to produce fitness differences, they must be in some way casually efficacious. Therefore, natural selection could not have been sensitive to normative truths. The only way in which natural selection would favor specific normative judgments because those judgments track mind-independent normative facts is by being somehow directed towards those facts in the first place. However, as already stated, this line of thought goes against the common view according to which natural selection is a blind mechanical process.

[84] Here it might be objected that the present argument rests on a mistaken construal of the tracking account. It might be claimed that the point of the tracking account is to explain the widespread presence of normative judgments with specific content. According to the tracking account, certain normative judgments are widespread because they are true and the ability to recognize those truths was advantageous. It is not claimed that the endorsement of certain normative judgments was advantageous because they were (or are) true. Even though I accept the objection to a certain extent, I think it is not well targeted. The reason for this is that Street (2006, p. 125) introduces the tracking account as an account of the relation between selective pressures and independent normative truths. Moreover, since the account is a *tracking* account it seems reasonable to construe it as claiming to explain the evolutionary advantageousness of certain normative judgments in terms of their ability to track the truth, that is, in terms of their being true. The explanation of the ubiquitousness of those judgments, on this construal, is reached as a consequence of the tracking account and not as the original explanandum.

Let us return to the first objection, according to which the tracking account is less plausible than the adaptive link account because it disregards the main function of normative judgments. Plausibly the main function of normative judgments is their motivational role in reinforcing adaptive behavior and, we can add, weakening maladaptive behavior. The most important identified candidate for the function of conscious normative judgments is overcoming the need for immediate gratification of desires in favor of satisfaction of long-term interests (see e.g. Ainslie, 2001; Joyce, 2006, chapter 5; Krebs, 2011; Rosenberg, 2011, chapter 6). This ability for normative governance (Gibbard, 1990) and self-regulation enables long-term cooperation and coordination between different individuals at different places and times by *motivating* appropriate behavior regardless of the capacity of the corresponding normative judgments to represent an independent order of normative facts. The tracking account construes the function of normative judgments in accordance with its capacity to represent an independent order of normative facts. However, as stated, from the evolutionary perspective *this* representational capacity can be only incidental and is therefore not a very plausible candidate for adaptation (Parfit, 2011b).

Let me elaborate on this argument. For the object-based theorist's conception of reasons to work in the context of evolutionary considerations, she needs to suppose that the evolved rational capacities somehow manage to recognize and respond to object-based reasons whose existence is prior and independent of those rational capacities (Hooker & Streumer, 2004). Determining rationality by prioritizing the notion of a substantive reason or a fact that counts in favor of something induces certain requirements on the form rationality can take in these accounts (see Korsgaard, 2011). That is, the order of determination forces us to see rationality as a kind of capacity for recognizing reasons akin to perception. Since reasons are facts that count in favor of something regardless of facts about the thinking agent, the task of the agent seems to be to *perceive* where the balance of reasons lies and use her rationality to respond to those facts by forming appropriate judgments or beliefs.

From an evolutionary perspective, the proposed hypothesis– namely that that the function of evaluative judgments is akin to perception – is totally implausible. The account proposes that the function of evaluative beliefs is to represent normative reality, just as perceptual judgments pertain to represent empirical reality. But in the normative case, as the discussion about the tracking account showed, the fitness benefits of having this additional ability to recognize the reality about normative reasons are not clear when those facts, unlike empirical facts, do not exert any causal influence on our behavior. According to the adaptive link account, the function of evaluative judgments is not to represent mind-independent normative reality but to reinforce

the behavior that at some point in our evolutionary history had some fitness benefits. In other words, the predicted function is *motivational*, and not cognitive (Gibbard, 1990; Joyce, 2006). The natural function of evaluative judgments is to motivate certain sorts of behaviors, not because they fit some facts about mind-independent reality, but because those behaviors and patterns of thought enabled organisms to cope with adaptive problems and selective pressures.

One might argue that on object-based accounts there is room for acknowledging that the capacity to deliver evaluative judgments encompasses both functions. In general, I don't think there is anything implausible about the idea that evaluative judgments play cognitive and motivational functions at the same time.[85] However, the point of the present argument is that this idea cannot be plausibly incorporated into an object-based account of reasons.

The reason why it is not available to the robust normative realist is because she needs to suppose that the primary function of evaluative beliefs is to represent mind-independent normative reality and that the motivational function is just a secondary or acquired function.[86] Because the cognitive function does not necessarily motivate, the normative realist would need to suppose that an evolved rule transforms those cognitive representations into action-guiding principles or motivations. For example, we would need to suppose that there is a rule of transformation according to which when a person judges that she has a sufficient reason to Φ or that she ought to Φ she, *ceteris paribus*, forms the intention to Φ. We can call this rule of transformation the *enkratic disposition*, after a similar principle proposed by Broome (2013, p. 13).

However, since the enkratic disposition could be favored by natural selection only if evaluative judgments are such that, at some point in human history, they reinforced fitness-benefitting behavior, then it follows that unless the primary function of evaluative judgments is motivational it is unlikely that the enkratic disposition would have evolved. If the primary function of evaluative judgments is to motivate fitness-benefitting action then it cannot be the case that what agents experience as counting in favor of and consequently judge that they have

[85] This idea became prominent among naturalistically minded philosophers. See e.g. Millikan (1996) and Harms (2004, chapter 8) for an account of representations that have both motivational and cognitive functions (see also chapter 5 of this thesis).

[86] This supposition needs to be made because if it were the case that cognitive and motivational functions were one and the same, if they emerged and were selected in the same period of time, then the idea of judging that there is a sufficient reason to do X and doing X would always be aligned. This is implausible because it would rule out *akrasia*, that is, the phenomenon according to which we can intentionally act against a judgment about what we have a sufficient reason to do.

a reason to do will generally reflect mind-independent normative reality. Rather this will reflect selective pressures that played a role in determining the organism's inclusive fitness. From these considerations, it follows that a normative realist who thinks that the primary function of evaluative judgments is to depict mind-independent normative reality cannot explain the existence of the mechanism that transforms those judgments into dispositions to act. Therefore, object-based theories of reasons cannot explain the emergence of the motivational function of evaluative judgments or explain why there should be a reliable connection between an evaluative judgment and the motivation to follow what those judgments recommend.

4.5 Response-dependence, common-sense, and evolutionary considerations

Let us take stock and overview the implications of the core argument of this chapter. The argument is that the object-based theorist has to decide whether to accept the premise that evolutionary processes (of which natural selection was the most significant) had no relation to mind-independent normative truths that we endorse or to accept the premise that there is some connection between mind-independent normative truths and evolutionary processes. The claim is that if the former horn is accepted one is left with an implausible form of normative skepticism. Furthermore, I argued that even if an account could be devised that explains how robust normative realism predicts the truthfulness of some normative judgments despite the influence of evolutionary forces, this move would not be in the naturalistic spirit. This is important because the argument is targeted against those who want to claim that robust normative realism is compatible with a broadly naturalistic world-view. Hence, the naturalistically sound move is to accept the second horn of the dilemma.

As we have seen, accepting this idea does not help the normative realist's case, because it puts her in a dialectical position where her preferred theory of the relation between the normative reality, capacities to grasp that reality, and evolutionary processes has to compete with other possible accounts of the same relations. This strategy is nicely illustrated by Street's (2006) discussion of the second horn of the dilemma. Street contrasts the two accounts and evaluates their merits in comparative terms, so that the failure of one is to the advantage of the other. However, even resolving the second horn of the dilemma in favor of the adaptive link account does not, by itself, warrant the ontological conclusion that object-based theories of reasons are false as a description of normative reality. The reason for this is that the adaptive link account by itself does not commit one to a specific ontology of normative facts, even though it is suggestive in that respect.

In fact, the comparative plausibility of the adaptive link account can be turned into an argument in favor of a certain type of subject-based theory of reasons. Subject-based theories fare better in accommodating evolutionary considerations into their own framework; their ontology is not robustly realist; rather, the supposition is that normative reasons reflect the agent's contingent nature. In particular, response-dependentist subject-based theories provide a nice unifying account of normative reasons that nicely fits into the evolutionary picture.[87]

Let me elaborate on this point a bit further. The argument of this chapter was targeted against the claim that the truth-conditions of normative claims are stance-independent. If we want to hold on to the two other intuitive claims, namely that normative judgments pertain to the truth and that some of them really are true, then a reinterpretation is in order. I propose that the reinterpretation be directed towards a dispositionalist or response-dependentist construal of the core claims about normative reasons. There is a plausible principle that instructs us when some concept should be interpreted as a concept of a response-dependent property.

> We, the theoreticians should introduce a response-dependent concept (or re-interpret an existing concept as response-dependent) only if we cannot identify a suitable categorical basis, or think there are principled reasons that prevent the thinker from referring to them. *(Miščević, 2006, p. 6)*

For example, the concept of a color seems to refer to intrinsic properties of objects. However, scientific data seems to show that there is no categorical basis with which colors could be identified (Giere, 2006; Hardin, 2003). Since we think that attributions of colors to objects can nevertheless be true and adequate, the latter principle instructs us to reinterpret the concept of a color as referring to response-dependent properties. Similarly, if we want to continue treating normative judgments as having truth-conditions, then in light of the evolutionary considerations the principle instructs us to construe the concept of a normative reason as response-dependent.[88]

---

[87] Dispositional or response-dependent properties are usually construed as having the following form: "The property F = the disposition to produce R in S under C, where R is the manifestation of the disposition, S is the locus of the manifestation, and C is the condition of manifestation" (Miščević, 2006, p. 5).

[88] For example, the response-dependentist view of reasons might be provided in the following equivalence: fact p is a reason R for agent A to do Φ in circumstances C iff in C, A *treats* p as a reason for doing Φ, where judging that r is a reason is something different from *treating* p as a reason. Obviously the relation of *treating something as a reason* would needed to be spelled out in more detail here in order to make the account of reasons more illuminating. As a first approximation we could say that *treating something as a reason* includes playing an appropriate inferential or functional role that ultimately has the potential to motivate the relevant behavior (Bratman, 1996).

Response-dependence explains the apparent fit between the intuitive pull of certain normative judgments and the maximization of inclusive fitness because it ties reasons to a person's responses, which of course have an evolutionary pedigree. In particular it can explain why we think that pain is bad, that there is a reason to take care of one's own children, to reciprocate, to endorse social cooperation, to avoid free-riding, to punish free-riders, etc. without presupposing a naturalistically suspicious normative ontology. The explanation goes *via* scientific accounts that try to explain how evolutionary 'design' problems introduced selective pressures for certain kinds of preference and evaluative profiles that in turn served as a basis for agents' rational capacities to work with. In this respect the account is clearer than the tracking account, because it does not misconstrue the mechanism of natural selection as being purposive. Rather it takes it as it is hypothesized in evolutionary theory and pertains to build normative claims on top of it. It is consistent with the predicted primary evolutionary function of normative judgments, that is, it incorporates the directive function into their truth conditions (see footnote 88).

Furthermore, it can explain why there could be a change in a person's reasons and in which direction the dynamics would go. For example, plausibly children have different reasons for action than adults, because they face different circumstances and problems that relate to their inclusive fitness maximization (similar considerations apply to people across different times and places). This account can also explain how people's reasons might be separated from their fitness considerations. Our rational abilities enable us to be flexible in our thinking processes and to imagine and contemplate ideas that do not necessary relate to biological fitness. The ability to think about the causes of our thoughts and actions enables us to take control over them, which plausibly implies having the ability to change them (Dennett, 2003).

However, wherever the dynamics of reasons takes one, truths about those reasons will not be mysterious because they will reflect the attitudes and abilities that that person possesses (as well as her social and natural environment) and not some everlasting normative reality. The bottom line is that a response-dependentist account can nicely accommodate the scientific insights and theories of the evolution and development of normative tendencies, judgments, and relations between people because it essentially construes reasons as functions of person's attitudes and responses to the circumstances in which she finds herself (cf. Street 2006).

4.6 Objectivist rejoinders

4.6.1 Third factor explanations and the pre-established harmony

An influential response to evolutionary considerations against robust normative realism is the claim that evolutionary processes, such as those that involve evolution by natural selection, are not, in a qualified sense, value-neutral (Behrends, 2013; Enoch, 2010; Parfit, 2011b; Skarsaune, 2011). The idea is that there is a pre-established harmony between the original results of natural selection and the evaluative attitudes that were formed as a consequence of selective pressures.[89] If we suppose that, for example, survival is at least somewhat valuable and that facts about survival provide one with object-based reasons to do and believe certain things, then it would be plausible to think that evaluative beliefs that were shaped by evolutionary processes are not completely off track with respect to independent normative facts. Were this proposal plausible we would have an explanation and dissolution of the Darwinian dilemma. There would be grounds for arguing that evolutionary processes pushed us in the right normative direction, because evolution by natural selection, which favors organisms that care about their survival prospects, shaped the contents of our normative judgments. We would be able to say that "[t]he fact that (roughly speaking) survival is good pre-establishes the harmony between the normative truths and our normative beliefs" (Enoch, 2010, p. 431).

However, this suggestion is totally implausible from the perspective of someone who endorses some version of the evolutionary debunking argument against robust normative realism (Behrends, 2013, p. 492). According to the evolutionary challenge, the idea that survival is in some sense independently valuable is a result of our evaluative beliefs being shaped by natural selection. Organisms that care about their survival or have a disposition to behave as if they care enhance their survival prospects and leave more progeny.[90] The ability to form judgments about the desirability of survival is plausibly explained by utilizing the same

[89] This idea of a pre-established harmony with respect to normative truths and evolutionary processes is also called a 'third factor explanation' (Behrends, 2013). The idea is that there is an additional (third) factor that breaks Street's dilemma and that might explain the alignment of evolutionary forces and the content of our normative judgments. In the most influential version of the third factor account, Enoch (2010) proposes that our survival's being *mind-independently valuable* provides a link that explains how natural selection might have aligned our normative judgments with mind-independent normative truth.

[90] Skarsaune (2011) gives a pre-established harmony account in terms of pleasures and pains that correspond to good and bad qualities, respectively. However, I believe that the same considerations given in the main text count against the plausibility of the *conditional strategy* that he provides in order to discharge the first horn of the Darwinian dilemma.

mechanism that enabled organisms to survive in the first place because it strengthened and reinforced the original connection between the relevant behavioral patterns and the probability of surviving and eventually reproducing. Claiming that survival is independently valuable and therefore that natural selection does not arbitrarily (with regards to normative reality) shape evaluative judgments is exactly what we would expect to believe if the evolutionary explanation was true. Because of this, the justification for thinking that facts about survival provide or represent some independent source of normative reasons is defeated.

Defenders of object-based reasons could complain that this objection makes too stringent dialectical demands on them. Parfit claims that this objection demands that its opponents defend their position without "making any assumptions about which normative beliefs are true," and that this is an implausible requirement "because we couldn't possibly show that natural selection had led us to form some true normative beliefs without making any assumptions about which normative beliefs are true" (Parfit, 2011b, p. 533). He illustrates the objection with an example:

> Some whimsical despot might require us to show that some clock is telling the correct time, without making any assumptions about the correct time. Though we couldn't meet this requirement, that wouldn't show that this clock is not telling the correct time. *(Parfit, 2011b, p. 533)*

However, the example misses its target, because Street's objection does not put normative realists in an impossible position. At least not in this concrete example, because the supposition that survival has some independent value is a direct target of the evolutionary argument and not some innocent background condition that dialectically lifts the realist's position off the ground.

Furthermore, it is not clear why would one even claim that survival is of independent value when we take into consideration what survival in its bare form consists of. Survival is just perpetuation of life, and life is just a certain form of organization that emerges from a combination and the recursive reactions of chemical elements (Griesemer & Szathmáry, 2009). From this perspective, we can say that bacteria and humans are both alive, and by surviving they continue to be alive. From the point of view of independent normative reality, it is not clear why we would value our own survival and not the survival of some other entity, such as a bacterium. However, from the perspective of the evolution of our own evaluative judgments it is clear why we would do so.

Behrends (2013) argues that the meta-normative realist can respond by arguing that according to his realistic theory we can expect to have a reason to take care of our own existence

and survival prospects because surviving provides the necessary means for doing almost anything else that might be of independent value. However, this maneuver does not work because even though the truth of robust meta-normative realism could explain why we think that survival has some instrumental value and how other things could become instrumentally valuable it does not take into consideration the fact that the Darwinian challenge calls the whole idea of robust normative realism into question. So the fact that conditionally robust realism entails the value of survival and our belief in it does not mean much when we have a reason to doubt the antecedent of the conditional.

Let me elaborate on this last point. My contention is that the evolutionary explanation of normative phenomena makes robustly realistic accounts of normative reasons superfluous. One of the implications of the discussion in section 4.4 is that evolutionary theory has enough resources to explain the emergence of many of our deeply entrenched normative beliefs without invoking their truth. This, in effect, makes accounts that presuppose mind-independent truths about basic normative reasons superfluous. Compare this to the following case. Modern evolutionary theory developed by Darwin and his successors explains how complex adaptations arise and why the appearance of design in nature is ubiquitous. The explanation of adaptation derived from evolutionary theory makes explanation of the same phenomena in terms of a supernatural designer superfluous. So the fact that the evolutionary explanation of adaptation can be made logically consistent with the existence of an intelligent designer[91] does not make that kind of explanation plausible. The evolutionary explanation makes the designer hypothesis obsolete because it does not add any explanatory value to the explanation that does not use that concept. Similarly, postulating mind-independent reality, to which our normative beliefs come to conform through evolutionary processes, does not add any value to the explanation of why we had those beliefs in the first place.

4.6.2 Normative beliefs and cultural evolution

Another possibility for a robust normative realist is to argue that the actual biological theory of evolution does not or cannot explain the emergence of intuitively held beliefs about normative reasons. However, there could still be a naturalistic explanation of the emergence of those

[91] For example, the designer could be someone who stands at the beginning of time and lets events unfold as they actually did.

beliefs that is more amiable to the account endorsed by the robust normative realist (or that at least does not present a threat to its consistency with a naturalistic worldview).

In section 119 of his (2011b), Parfit points to some issues that are rarely explicitly discussed in the meta-ethical literature. It is worth quoting Parfit at some length:

> We can first note, that, when Street and others make claims about the effects of evolutionary forces, these writers are not referring only to genetic evolution. Just as certain genes became more widespread when people with these genes were more likely to survive and pass on these genes to their children, certain beliefs became more widespread when communities of people with these beliefs were more likely to be successful, in ways that preserved and spread these beliefs. So we should ask which normative beliefs would have been advantageous either reproductively, or at the social or cultural level. *(Parfit, 2011b, p. 534)*

A few pages below he continues this line of thought:

> Some normative beliefs became more widely spread when and because communities of people with these beliefs were more likely to be successful. It is much less clear how we should assess the claim that certain normative beliefs were in this way, not reproductively, but socially or culturally advantageous. It is less clear, for example, whether and how such explanations of our normative beliefs should be assumed to debunk or undermine these beliefs. When the acceptance of certain normative beliefs made some community or culture more likely to survive and flourish, this fact does not as such cast doubt on the truth or plausibility of these beliefs. Such explanations of our normative beliefs do not obviously, in Street's phrase, contaminate these beliefs. *(Parfit, 2011b, p. 537)*

In these two passages, Parfit points to an important line of research that should be pursued in order to see which normative beliefs might be more plausibly explained in terms of cultural group selection (Richerson & Boyd, 2005). Parfit points out that often the widespread acceptance of many normative beliefs cannot be easily explained in terms of their impact on genetic fitness. For example, he mentions our acceptance of the wrongness of lying, breaking promises, stealing, and the acceptance of some versions of the golden rule (Parfit, 2011b, pp. 536-537) as examples that do not seem to be easily explained in terms of the benefits they have regarding the proliferation of our genes. Rather they might be more easily explained in terms of selection that acts at the level of cultural groups. For example, the golden rule might be favored at the level of the group, since it promotes cooperation and because groups that cooperate fare better than those that do not.

However, even if we admit that evolutionary processes at the level of cultural groups had important effects on the formation of our normative beliefs, this does not mitigate the force of the evolutionary debunking argument. There are at least three interrelated reasons for

reinforcing the Darwinian challenge at the level of cultural group selection. One is that cultural evolution cannot make some trait adaptive if that trait is maladaptive at the biological level (André & Morin, 2011). In this respect, formal analysis of the relation between the evolution of genes and culture has shown that genetic selection constrains which cultural items (beliefs, behaviors, norms, institutions, etc.) will be favored by natural selection (El Mouden, André, Morin, & Netlle, 2013). Second, and more specifically, evolved human cognitive biases constrain and tend to eliminate cultural traits that are maladaptive (ibid.). In this respect we can say that in the long run only those normative beliefs that are advantageous or neutral with respect to inclusive fitness maximization will survive. This consideration reinforces the problem of the pre-established harmony that a normative realist needs to explain.[92] Third, even if we abstract from the explicit relation between genes and culture, it is widely recognized that cultural formation and transmission of beliefs often leads to the wide proliferation of false beliefs (Boudry, Blancke, & Pigliucci, 2015). So just because some trait might be culturally advantageous this does not mean that it must accurately reflect the structure of reality. For example, it is widely recognized that religious beliefs might have a group-level function in that they have the capacity to homogenize and stabilize a group of people (Boyer, 2001). However, this possible social function of religious beliefs does not in any way indicate that they are grounded in mind-independent facts.

4.6.3 Do cognitive explanations of normative beliefs trump evolutionary explanations?

Ultimately, Parfit also recognizes that cultural evolutionary explanations will not help the normative realist in dissolving the Darwinian dilemma (Parfit, 2011b, p. 538). Nevertheless, he rejects the dilemma on the basis of the contention that neither genetic nor cultural evolution can

[92] Someone might claim that the result according to which natural selection will weed out normative beliefs that are maladaptive in some way (such as that we have a reason to run off a cliff) might be interpreted as giving support to the claim that there is a pre-established harmony between the basic goals of fitness maximization and the basic value of survival and reproduction. However, we would need to see how that argument might be developed in order to seriously discuss it. Second, it seems to me that it is more natural to suppose that if natural selection, in the long run, circumscribes the space of what is normatively possible then whatever is contained within must serve as constructive grounds for producing normative beliefs, and not as capturing something that is already there. Again, the thought behind this is that it is not clear why some contingent space of possibility should reflect mind-independent truths about the normative domain.

explain the formation of our intuitively normative beliefs. His contention is based on the following considerations:

1) If we can devise evolutionary explanations for the formation of two or more *conflicting* normative beliefs then this fact will count against those explanations if there some other plausible alternative explanation of those beliefs is available.
2) We can often devise evolutionary explanations for the formation of two or more conflicting normative beliefs.
3) Normative beliefs can be plausibly explained in other non-evolutionary based ways.

If we construe the statement 'that fact will count against those explanations' as a prescription to search for other plausible alternative explanations we can conclude:

Therefore,

4) we should adopt those other kinds of explanations for the formation of our normative beliefs.

This argument is a rough formalization of the considerations provided by Parfit (2011b, p. 536). Parfit supports 1) by giving examples. For example, he seems to say that we could devise an evolutionary explanation of why raping and committing adultery is believed to be wrong. But similarly, if we believed that men ought to rape women and commit adultery, we could also explain why that would be the case in terms related to genetic fitness (Parfit, 2011b, p. 535). For the case of cultural evolution, we can use the example of the Golden Rule. If everybody followed the Golden Rule then everybody would reap the benefits of cooperation that is supported by such a rule. However, cultural group evolution might also have favored the production of normative beliefs that give priority to our own group-members at the expense of all other people that belong to other groups, and in that sense it could have "helped some communities to destroy, conquer, or exploit others" (ibid., p. 537).

I am not sure how seriously we can take Parfit's requirement that evolutionary explanations do not apply if they have the capacity to provide explanations of different contingent paths that the history of the organic and cultural world might have taken. After all, evolutionary processes are highly contingent (Beatty, 1995), such that history could have played out differently and we could have evolved radically different traits and abilities. To assess the plausibility of Parfit's requirement we would have to go deeper into what philosophy of science has to say about the structure and plausible requirements on explanations in this context. This is a task for another paper. Here I will concentrate on the other condition that Parfit introduces in his requirement,

namely the idea that in the context of the formation of normative beliefs we have a better alternative explanation at our disposal. I will argue that Parfit's argument fails because his proposed explanation does not provide a *genuine* alternative to evolutionary explanations of the underpinnings of our normative beliefs. In other words, even if the requirement applies in this case it is not violated because we do not have other genuinely non-evolutionary explanations of how our normative beliefs were shaped and ultimately formed.

There are three salient types of explanation of the formation of our deeply held normative beliefs that Parfit seems to take as possible alternatives: biological evolution, cultural evolution, and individual cognition. According to Parfit, our ability to respond to reasons and the intrinsic credibility (i.e. their intuitive appeal) of normative contents better explains the formation of normative beliefs then their possible fitness value (Parfit, 2011b, pp. 535-536). Therefore, Parfit opts for the third type of explanation.

However, this suggestion is not a genuine alternative to evolutionary accounts because it confuses the proximal/distal distinction. Evolutionary accounts provide distal explanations of the emergence of traits, their maintenance in the population over long periods of time, and even their functions, i.e. they pertain to explaining why some trait has evolved. Accounts that explain the emergence or promulgation of some trait in terms of our cognitive capacities rely on proximal mechanisms (attention, memory, decision-making, reasoning, etc.) to explain how that trait became common or why it was adopted. So Parfit's purported explanation of how our normative beliefs were formed does not compete with evolutionary-based explanations, they just operate on different levels.

Let me substantiate this claim. The standard view is that evolutionary processes shaped our mechanisms for acquiring beliefs; they shape the inputs on which those mechanisms work, they shape their operations, output conditions, and the way in which they operate in concert or in isolation from other mechanisms (see section 4.4.3 above, and Street (2006)). Hence, just proposing that cognitive abilities explain the formation of normative beliefs does not answer the Darwinian challenge. The problem is not whether cognitive abilities or evolutionary forces directly produce normative beliefs. Rather the problem is determining the ultimate grounds of those beliefs. The robust normative realist wants to claim that those grounds lie in mind-independent normative reality, while the arguments provided in this chapter suggest that those grounds lie in facts picked out by mechanisms that were shaped by evolutionary processes.

The bottom line of this argument is that if someone wants to claim that our cognitive abilities can, in an autonomous way, explain how our normative beliefs, whose actual function is to solve practical problems (e.g. feeding, surviving, mating, reproducing, etc.) encountered

by humans in their evolutionary past and in their present circumstances, were formed, then she needs to explain how those abilities come to be 'unshackled' from the 'chains' of their evolutionary origins and for what purpose. I predict that this task will be difficult to execute for the following reasons: the cognitive capacities underpinning decision-making are most centrally related to fitness (Sober & Wilson, 1998, p. 159), and therefore our decision-making capacities were most likely to a large extent shaped by biological and cultural selection forces. This in effect shaped our more reflective normative judgments (see section 4.4.1, whose function is (in one way or another) to reinforce behaviors that ultimately work in favor of maximizing our inclusive fitness (El Mouden, Burton-Chellew, Gardner, & West, 2012).[93] Earlier I remarked that even cultural selection would not be able to run its course without the constraints of selection acting at the level of genes (El Mouden, André, Morin, & Netlle, 2013). However, this applies even more to our cognitive abilities because they have a more direct influence on our biological fitness.

It could be argued that the process of gene-culture coevolution[94] can explain how our cognitive abilities were able to become independent from their genetic underpinnings and even to create selection pressures for genes that 'work in their favor.' For example, early humans developed cognitive abilities that promoted their inclusive fitness. Then those abilities served as underpinnings that promoted the development of cultural artifacts, skills, etc., which in turn served as the basis of further development of our cognitive abilities. An example is the development of formal teaching and the transmission of knowledge thorough horizontal and vertical chains, which further enabled the development of cognitive skills and capacities. Finally, at our present stage of development and knowledge we are able to control (to a certain extent) our environment, including the selection pressures that act on our genes. This idea

[93] This does not mean that we do not often and recurrently fail to act as if we are designed to maximize our inclusive fitness. For the review of the idea that people evolved to act as if they maximize inclusive fitness and the reasons for which we often fail to act in this way, see (El Mouden, Burton-Chellew, Gardner, & West, 2012).

[94] According to the gene-culture coevolution view, genes and culture form two separate but interactive systems of inheritance "with offspring acquiring both a genetic and a cultural legacy from their parents and, in the latter case, other conspecifics too" (Laland, 2008, p. 3578). The most important message of the gene-culture coevolution view is that culture is an important source of genetic evolution, in the sense that culture and its mechanisms of transmission can modify the evolutionary environment, which in effect can modify the selective pressures that act on genes. The most widely cited example of this phenomenon is the coevolution of lactose absorption and human dairy farming (ibid.). It is widely believed that dairy farming spread before the gene for lactose absorption. Consequently, farming provided selection pressures for genes for lactose absorption to spread in the population of early dairy farmers.

regarding the way in which the development of our cognitive abilities might have allowed us to come into control of survival and reproductive opportunities could also be used to support the robustly normative accounts that rely on third-factor explanations (see section 4.6.1 above).[95]

The presupposition of this account is that at some point in the human history there was a group of people who discovered real mind-independent truths about many normative reasons. They then managed to persuade enough other people to believe in these truths, and thereby to modify the cultural environment in such a way that selective pressures would be neutral or beneficial towards the fixation of those beliefs. This is obviously a very simplistic exposition of the idea under examination. Nevertheless, something along these lines must have been the case if we take seriously the possibility that cognitive abilities managed to restrain the otherwise contingent path of gene-culture coevolution and led to the truths posited by robust normative realists.

However, this proposal cannot help the robust normative realist to make her case. The general problem is the non-parsimonious assumption about the components that need to be posited in order to provide a plausible explanation of the formation of many of our normative beliefs. A plausible explanation of how our normative beliefs were detached from the obvious tendencies to maximize inclusive fitness goes roughly along the following lines. The natural tendency of any well-adapted organism (including people) is to maximize its inclusive fitness. However, organisms generally do not consciously maximize their inclusive fitness. Rather, natural selection favored the evolution of those organisms that managed to react and respond to cues that were reliably related to fitness. In the case of human beings, those proximal mechanisms that evolved as responses to fitness-related reliable cues can be broadly classified under the terms 'pleasure' and 'pain' (El Mouden, Burton-Chellew, Gardner, & West, 2012). For example, the pleasure that was provided by engaging in sexual activity served as a reliable cue to reproductive success. The fact that fire burns served as a cue to stay away from it, etc. The evolution of positive and negative affects provided a base for other more sophisticated emotions and cognitive abilities. However, since in this simple model pleasure and pain play the most basic elements that serve as cues for fitness-related opportunities, as soon as people became somewhat self-conscious they envisaged those affective states as the most elementary

[95] For example, an explanation of the pre-established harmony between our normative beliefs and the 'goals' of natural selection might be that our normative beliefs about the instrumental value of survival caused further alignment between what we value as products of evolution and what we objectively have a reason to want or do.

components that ground their actions. We can easily imagine the evolution of proto-normative states whose content is 'pleasure is good', 'pain is bad,' 'more pleasure is better than less', 'less pain is better than more', etc. In other words, when capacities for thinking and valuing evolve they naturally tend towards valuing pleasure intrinsically and believing that pleasure (or a person's well-being more generally) constitutes intrinsic value.

We can then imagine how people's desire sets or sets of considerations that psychologically relate to their well-being got detached from the direct tendency to act as if they were maximizing their inclusive fitness.[96] Since people's basic motivational and evaluative outlook is at root grounded in the attainment of pleasure, the emergence of greater cognitive abilities allowed them to seek and find better and more efficient ways to attain pleasurable things or engage in activities that provided them with pleasure. For example, greater cognitive abilities enabled people to invent birth control and to have safe sex without worrying about accidental pregnancies and thus having to allocate resources to taking care of unplanned progeny. Incidentally, the invention of ways of having safe sex might have lowered people's inclusive fitness. However, with the invention of condoms and other wonders of the modern world the ecological niche in which people function changed so radically that even selection pressure towards increasing one's reproductive success diminished.

This is a very simplified model of how greater cognitive abilities might have acted on the basic traits that were proximal indicators of fitness value and detached them from their original role (direct inclusive fitness maximization). Presumably the normative realist wants to tell this kind of story when she claims that gene-culture coevolution and greater cognitive capacities enabled people to reach mind-independently true normative beliefs. However, if this is the model endorsed by the robust normative realist then she is introducing additional assumptions for which there is no obvious scientific justification. The realist wants to say that on top of everything that is described in the simple model, the fact that a person *takes* something to be intrinsically pleasurable and therefore starts to value it intrinsically means that we need to add additional ingredient, namely that the things that are valued intrinsically *have* a further property

[96] See Sterelny (2012) for a more sophisticated account of how the explanation of human agency benefits from modeling it in different ways in relation to its development through evolution. For example, in early human history when the group cohesion was greater and the fitness-related transmission of information was vertical (parent to offspring), human agency could be modeled as if it more directly aimed at inclusive-fitness maximization. However, when social groups become larger and more diverse, and information is transmitted horizontally, it becomes more plausible to model human agency as if maximizing expected utility, where, of course, utility and inclusive fitness represent different quantities that may be more or less connected.

of actually being mind-independently valuable (or reason-providing). It is hard to find an independent justification for this further ontological assumption.

Of course, what propels philosophers to think that they have some kind of justification is the *intuition* that certain things have intrinsically valuable or reason-giving properties. But once we recognize that those intuitions are probably grounded in the same mechanisms that at some point in our evolutionary history functioned as proximal cues to fitness-related considerations (pleasure and pain), they lose their evidential force. That is, they lose their force as evidence in favor of thinking that what we value intrinsically must also refer to mind-independent normative reality. From the explanatory perspective, everything about our basic normative beliefs is accounted for in terms of our attitudes, such as *valuing intrinsically* and *taking something to be a reason*. The additional claim that those things have actual intrinsic value (or reason-providing mind-independent properties) does not strictly play any role in the explanation of how our basic normative beliefs were formed or the fact that we may have evolved to detach ourselves from valuing direct fitness-relevant considerations. The bottom line of this argument is that even if the hypothesis that our cognitive abilities, through some process of gene-culture coevolution, took control of our biological nature and constructed a new environmental niche were plausible, still the hypothesis that our normative beliefs reflect mind-independent, abstract and non-causal normative reality would be scientifically implausible. That is, they plausibly lack explanatory parsimony. The conclusion of the argument points back to the themes of section 4.5. To clarify, my claim is that evolutionary explanation (even the gene-culture coevolution hypothesis) makes superfluous the hypothesis that normative beliefs refer (if they refer at all) to mind-independent, intrinsically valuable properties in the same way in which natural selection makes superfluous intelligent-design explanations of apparent design in nature.

4.7 Concluding remarks

In this chapter I have argued that the mind-independence thesis of robust normative realism about reasons does not dovetail very well with naturalistic theories that stem from evolutionary approaches to the mind. In so doing, I mostly relied on Street's (2006) argument based on evolutionary considerations. If a normative realist wants to deny that evolutionary processes are responsible for the truth-status of contents of normative judgments, then a thoroughgoing skepticism about our normative reasons would ensue. Furthermore, I argued that the normative realist who goes for the first horn of the dilemma in a way denies the possibility of a naturalistic

explanation of the normative phenomena examined in this chapter. In this respect, the argument of this chapter would probably not apply to the views of those authors. In that case, however, the normative realist would have to be deemed a scientific skeptic. If the normative realist accepts that evolutionary processes stand in some relation to the mind-independent truth of normative judgments, she faces a further problem: namely, her explanation of how our evolved normative judgments track normative truth must be evaluated in comparison to other plausible evolutionary explanations of the adaptive value of those same normative judgments. In this respect, I argued that the normative realist's account (which Street (2006) calls the tracking account) is less epistemically virtuous than another non-truth-based evolutionary explanation (which Street (2006) calls the adaptive link account).

The bottom line of the argument is that strictly speaking the mind-independent truth of normative judgments does not play a role in the explanation of why those judgments played an important fitness-enhancing role in our evolutionary history. However, I further argued that if we want to hold on to the claim that our deeply held normative judgments about reasons express some truths then they should be minimally reconstrued as judgments about response- or mind-dependent properties. This position dovetails more neatly with evolutionary considerations since the mind and its behavioral outputs, according to the current scientific orthodoxy, also have an evolutionary history and functions that were shaped by it.

5 The emergence of reasons and rationality

5.1 Introduction

The aim of this chapter is to develop an account of normative reasons that respects the constraints provided by the evolutionary argument discussed in the previous chapter. This includes providing an account of normative reasons that construes them as mind- or attitude-dependent entities.

A plausible theory of normative reasons should satisfy at least two desiderata. First, it should explain the existence of reasons that people have in virtue of being agents that have desires, goals, or aims. In Kantian terminology, those would be *hypothetical* reasons. Second, the theory should explain why we experience some reasons as transcending particular occurrent desires, goals, or aims. In Kantian terminology, those would be *categorical reasons.* The differentiation between these two types of reasons is phenomenological.[97] Hypothetical reasons seem to be such that their normative force depends on our having certain attitudes. Categorical reasons phenomenologically seem to be those whose normative force does not depend on our having particular goals or aims.

In this chapter, I will argue that a subject-based theory of reasons can account for the difference between the two types of reasons. I will try to show this by developing a naturalistic story about how reasons could have emerged and become fixed by responses of agents belonging to different levels of cognitive complexity. In developing such a story, I will rely on the supposition that the concept of rationality might provide the foundations for identifying the sources of our practical reasons. From this discussion, it will emerge that the distinction between hypothetical and categorical reasons depends on the type of rational principles we adopt. The principle of instrumental rationality will account for hypothetical reasons. More substantive principles need to be presupposed for determining categorical reasons. However, traditionally naturalists have had a difficult time explaining how we could adopt principles that go beyond the instrumental principle of rationality. In order to account for this possibility I will rely on a game-theoretic model that explains how primitive semantic relations get established

[97] Of course, it should be added that some normative realists see this phenomenological difference as indicating an ontological difference between types of reasons.

among a community of agents. I will contend that the same model can be applied to naturalistically explain how phenomenologically categorical reasons could emerge from beneficial interactions among different agents.

This chapter is divided in the following way. In section 5.2 I will explain the difference between hypothetical and categorical reasons. In section 5.3 I will discuss the relation between three fundamental concepts: the faculty of reason, rationality, and substantive reasons. I will adopt a view according to which the faculty of reason and the principles of its functioning determine what substantive reasons we have. This view will be justified from a naturalistic viewpoint because we tend to apply different criteria of rationality depending on agents' levels of cognitive and behavioral complexity. In section 5.4, I will introduce principles that might individuate hypothetical and categorical reasons. Then, by relying on a game-theoretical model of how primitive semantic relations are established, I argue that the same framework can provide a model for how categorical reason relations could emerge.

5.2 Hypothetical and categorical reasons

As mentioned, it seems that we intuitively distinguish between at least two types of reasons: hypothetical and categorical. Thus, an appropriate theory of reasons should be able to differentiate between these two types of reasons or, if it cannot, then it should explain why this distinction, against appearances, does not hold.

In common terms, hypothetical reasons are those reasons that essentially depend on an agent's desires, broadly construed. 'Essentially' here means that reasons depend on a particular agent's motivational set and its particular elements: if a desire is a part of the set then this provides a reason for satisfying it; if there is no desire then we lack such a reason. To illustrate this common idea, Jonas Olson gives an example:

> [T]here is a reason for me to visit the local bar this evening because they are showing a football match I desire not to miss. So the fact that the local bar is showing the match is reason for me to go there. But it is obvious that this fact's being a reason for me to go there is contingent on my desire not to miss the match. Were I somehow to lose my desire not to miss the match, the fact that it is shown at the local bar would, ceteris paribus, no longer be a reason for me to go there. In other words, I could escape the reason to visit the local bar this evening by dropping my desire not to miss the match. […] this indicates that my reason to visit the bar is hypothetical […]. *(Olson, 2014, p. 118)*

Categorical reasons, on the other hand, do not depend contingently on the particular desires of the agent. A paradigmatic example of how we philosophically think about categorical reason is provided by the reasons stemming from moral requirements. Once again, an example given by Olson can illustrate the difference:

> Suppose for instance that it is morally wrong to eat meat and that one ought morally to donate 10% of one's income to Oxfam. The fact that it is morally wrong to eat meat entails that there is a reason not to eat meat. The reason – the fact that counts in favour of not eating meat, that is – might be that eating meat is detrimental to human and non-human well-being. Likewise, the fact that one ought morally to donate 10% of one's income to Oxfam entails that there is a reason to do so. The reason might be the fact that donating to Oxfam promotes human well-being.
>
> In these cases the reasons are not contingent on the agents' desires. Whether or not agents desire to promote human and non-human well-being, they have moral reasons not to eat meat and to donate 10% of their income to Oxfam. […] One cannot escape moral reasons by adverting to one's desires in the way I can escape my reason to visit the local bar this evening by jettisoning my desire to watch the match. *(Olson, 2014, pp. 118-119)*

Categorical reasons, such as moral reasons, have a sort of *inescapability* that hypothetical reasons lack; it seems that they cannot be dismissed just by losing a desire to obey them. In addition to categoricity and inescapability, some authors claim that moral reasons, in particular, have an *(overriding) authority*, in the sense that when they come into conflict with other non-moral reasons, they tend to trump them (Brink, 1997; Cuneo, 2007; Joyce, 2006).

If one adopts a subject-based theory of reasons, then it seems that accounting for hypothetical reasons is not a problem. After all, on subject-based theories, reasons are provided by facts about an agent's desire, goals, concerns, etc. Categorical reasons, however, might pose a problem, since they should apply to agents despite their not depending on the agent's contingent aims, concerns, etc. Nevertheless, I will argue that categorical reasons could be conceived as a contingent extension of an agent's hypothetical reasons; they are subjective reasons writ large, so to speak. I will further claim that categorical reasons emerge through interactions between different agents, and thus could be construed as hypothetical reasons that emerge from a population of agents and apply to agents in virtue of their belonging to a particularly structured population.

5.3 Rational faculties and reasons

Christine Korsgaard (2009; 2011) distinguishes between three fundamental concepts in normative philosophy: the faculty of reason, rationality, and substantive reasons. Reason as a faculty is usually conceived as an active part of the mind that phenomenologically has a particular authority over our thoughts and actions – that thing that makes us uniquely human. In this context, rationality is plausibly construed as a set of principles that describe the proper activity of the faculty of reason. Finally, substantive reasons are the particular things, facts, or state of affairs that count in favor of something, that is, those things to which the faculty of reason responds.

Different authors construe the relation between these three concepts differently. As we saw in chapter 2 of this thesis, Parfit (2011a; 2011b) and other authors who think that normative reasons are irreducibly normative seem to place emphasis on substantive reasons and tend to explain rational capacities in terms of them. Other influential authors think that rational requirements are one thing and substantive reasons are something different, where neither is plausibly explained in terms of the other (Broome, 2013). Others still, such as Korsgaard (2011), think that the faculty of reason represents the basic source of normativity and that the nature of substantive reasons can be explained in terms of this. In this respect, I will follow Korsgaard.

One reason for this is that the alternative, namely that substantive reasons are something completely different, seems to be implausible to me. First, views that want to explain rationality in terms of substantive reasons can be reduced either to views according to which substantive reasons can be explained in terms of rational requirements, or to views claiming that rational requirements are one thing and substantive reasons are something else. Regarding the first disjunct, I will just point out that intuitions about what we have a reason to do can be interpreted as intuitions about how rationality requires agents to form beliefs and desires when they deliberate about what to do (see Smith, 2009, see also section 5.4 below). If one does not like the idea that intuitions about substantive reasons can be interpreted as intuitions about what rationality requires, that is probably because one is overwhelmed by intuitions such as those underlying the Williams' gin-and-tonic example (cf. Broome, 2007, p. 167). The intuition that people seem to have is not just that Mary is rational when she drinks the petroleum while thinking that it is gin and tonic, but that she would be irrational if she did not drink it, despite the fact that she does not have any objective reason to drink the petroleum. If one is persuaded

by this intuition, then one is probably prone to thinking that rationality could require something that does not have anything to do with what one actually has reasons to do.

The reason why I am not prone to adopting this view is related to how authors who do adopt this view tend to construe the concept of reason. For example, John Broome construes reasons as certain type of explanation of ought-facts. By way of illustration, let us consider Broome's definition of what he calls *pro toto* reasons: "A pro toto *reason* for *N* to *F* is an explanation of why *N* ought to *F*" (Broome, 2013, p. 50). According to this view, a normative reason is a fact that makes it that something ought to be the case, just as natural selection makes it the case that evolution occurs (ibid., p. 48).

The problem I see with this proposal is that it does not properly capture the role of reasons in deliberation and it seduces us into thinking that the outputs of reason relations involve judgments that refer to some self-standing ought-facts.[98] For example, when I think that I have a conclusive reason to believe that p I do not necessarily come to believe that those reasons provide an explanation for why I ought to believe that p. First, by thinking that I have normative reasons to believe that p, I could just believe that according to epistemic norm E I am justified in judging that p is the case. Alternatively, I could just think that the premisses that led me to the conclusion actually entail the conclusion, without thinking that those premisses really make it the case that I ought to believe the conclusion in some way that is external to the deliberative processes that led me to the conclusion.

Second, in the extreme case, I might not believe that there is anything I really ought to do or believe. However, even in this extreme situation, I could still think that there are better and worse reasons for believing things, and better and worse ways of doing things. It seems to me that there would still be some normativity that would need to be explained. For example, even if there were no purely normative facts about what to do or believe, we would still be confronted with tasks that we need to solve and decisions that we need to make. The conclusions that we would reach would often involve the idea that we *should* do something. However, this judgment about what we should do will be of a practical kind, and not something that represents a fact that we should try to explain when deliberating. If that judgment could be characterized as true, then it would be true because of something about the process that led to it. Of course, that

[98] My thoughts on the issue should not be construed as providing conclusive arguments against Broome's notion of a normative reason. Broome develops an important and in many ways subtle account of reasons and rationality and their relation to other normative concepts. Thus, the following considerations should just indicate why I personally do not prefer this way of thinking about normative reasons in general.

process would be characterized as normative, but the normativity would be of the kind that we normally connect with the rationality of deliberation and the tasks we are disposed to perform. Thus, in Korsgaard's words we naturally come to the view that "if reasons did not exist, we would have to invent them" (Korsgaard, 2011, p. 6). We would need to invent them to play a practical role in directing action. From this perspective, it seems clear that positing reasons as theoretical entities in some detached normative realm does not add anything to the practical role that reasons play in our mental economy.

In addition, Korsgaard's construal of the situation has naturalistic credentials. It provides us with handles that can be interfaced with concepts from the cognitive and evolutionary-based sciences. Let me illustrate the rough idea. On this view, substantive reasons do not come out as something strange and ontologically irreducible because reasons can be construed as things that provide inputs to the faculty of reason, and what they count in favor of is what the faculty of reason (when functioning properly) provides as outputs. Thus, the focus is put on the faculty of reason and its principles of rational functioning. Now the question is how to conceive those principles and how they can explain, in a naturalistically respectable way, the difference between hypothetical and categorical reasons. We can begin to answer this question by thinking about the function of the faculty of reason and its principles.

5.3.1 Levels and functions of rationality

In general, we can say that the role of reason or rationality is to enable a living being to successfully perform some task (Simon, 1956). Besides, in involving tasks, the concept of rationality seems to apply most naturally to situations in which an organism is faced with a 'space of alternatives' from which it can choose types or tokens of behaviors, so to speak (Bermúdez, 2003, p. 117). In the basic case, the task of every living creature is to live long enough to reproduce. Depending on the task that a creature is performing, different types of rationality evolved as an enabling condition to successfully preform the task. José Bermúdez (2003; see also Kacelnik, 2006) helpfully distinguishes between three types of rationality (or as we might say, three faculties of reason) that we can ascribe to creatures.

At the most basic level we find what Bermúdez (2003, p. 116) calls level 0 rationality. This type of rationality is basic in the sense that it involves the ability to form and learn adaptive responses in relation to fitness-relevant circumstances. This type of basic rationality is, for instance, involved in learning through simple classical or instrumental conditioning, which is already present in simple creatures such as fruit flies (Brembs, 2009). According to Bermúdez

(2003, p. 117), the application of the concept of level 0 rationality is "not grounded in any process of decision-making"; rather it applies to an organism's behavioral dispositions or the types of behaviors it is able to perform. In that sense, when we evaluate an organism's level 0 rationality we do not ask whether any particular action was appropriate or reasonable in relation to some goal – since it is not necessary for any real decision-making to be involved – rather we evaluate patterns or programs (algorithms) that determine behavior and that an organism is disposed to execute. These types of behavior can be implemented at the level of genetically based hard-wired behavioral procedures, but not necessarily, since they can involve domain-general learning systems such as classical and operant conditioning.

Thus, at this level of rationality, which even fruit flies can satisfy, behavioral dispositions will be evaluated in terms of short-term and long-term criteria. Among the latter, Bermúdez (2003, p. 118), following Dawkins (1986), includes the organism's general effort to maximize its inclusive fitness. The former criteria include satisfying more proximal goals, such as maximization of energy intake, making trade-offs between exploratory and exploitative efforts when foraging, balancing particular activities (such as mating and avoiding predators) related to reproduction and survival, etc. As, we will see, all other levels of rationality will include similar short-term and long-term criteria of evaluation.

At the top of the conceptual hierarchy of rationality is what Bermúdez (2003, p. 123) calls level 2 rationality. This is the full blown, common-sense concept of rationality that includes a sophisticated representational apparatus, a theory of mind, and the possibility of combining different mental states in decision-making processes. Here rational evaluations, in effect, apply to both particular actions (not just types of behaviors) and decision-making processes. In between levels 0 and 2 there is level 1, which is unlike level 2 since it does not include a sophisticated representational apparatus or decision-making, and unlike level 0 because it allows us to apply rational standards to token behaviors or actions. This level of rationality is important in the present context because it already involves a familiar sort of normativity. To see this let us examine how Bermúdez conceives of it.

The essential feature of creatures with level 1 rationality is that they confront the world (environment) as partitioned into opportunities for action, from which they can select alternatives in accordance with their prefixed needs or goals, but without engaging in any substantive or folk-psychologically familiar decision-making. To illustrate this idea, Bermúdez provides an example:

> Imagine an animal confronted with another potentially threatening animal. The animal has two possible courses of action – fight or flee. There is a clear sense in which one of the two courses of action could be more rational than the other. Roughly speaking, it will be in the animal's best interests either to fight or to flee. And it seems that in such a situation there need be no process of decision-making. The animal might just 'see' that fighting is the appropriate response. Or it might just 'see' that fleeing is appropriate. *(Bermúdez, 2003, p. 121)*

Here Bermúdez, following Gibson (1979), invokes the concept of *affordances*. This concept accounts for a form of direct perception that can be used to explain behavior without supposing that the organism produces action through some cognitively sophisticated decision-making. The concept of affordances enables us to see that perception does not just consist in sensing objective spatial and temporal relations in the environment; rather:

> [i]t involves seeing our own possibilities for action – seeing the possibilities that are 'afforded' by the environment. If this is right then we can see how a given behavior might be selected from a range of alternatives in a way that does not involve a process of decision-making. The comparison of affordances does not require a process of decision-making. Nonetheless it is assessable according to criteria of rationality. *(Bermúdez, 2003, p. 121)*

At this level of rationality, the concept of affordances enables us to interface normative reasons with a naturalistically respectable notion. It helps us to unpack the response part of dispositionalist accounts of reasons. We might say that even though affordances as possibilities of actions are objective, which action possibilities are relevant is still determined by the abilities, needs, and tasks that an organism has evolved to perform. According to James Gibson, affordances are relative to individuals (cf. Gibson, 1979, p. 128). For example, a child perceives a tiny chair, in Gibson's words, as sit-on-able, while an adult, being too tall for the chair, does not. In this sense, the *relevance* of the affordances provided by an environment is determined by the responses the organism is prone to making and the benefits it thereby acquires.[99]

In particular, affordances can account for the fact that the world is given to us as normatively painted. At the phenomenological level, we see things and situations as affording

[99] For the purposes of the analogy that I am trying to draw, it is important to stress that Gibson does not construe affordances as completely objective properties of environments. This is clear from the following quote: "An affordance cuts across the dichotomy of subjective-objective […]. It is equally a fact of the environment and a fact of behavior. It is both physical and psychical, yet neither. An affordance points both ways, to the environment and to the observer" (Gibson, 1979, p. 129). Thus, affordances can be naturally interpreted as response-dependent properties. This is the sense in which I think the notion of affordances can be used to illuminate the fact that to us the world is normatively given.

us opportunities for action or in more familiar terms, as *counting in favor of* doing something as opposed to something else (Street, 2006). In fact, when discussing the origin of reason Korsgaard describes the situation in similar terms:

> A nonhuman animal is guided through her environment by means of her perceptions and her desires and aversions: that is, by her instinctive responses and the other desires and aversions she may have acquired through learning and experience. Her perceptions constitute her representation of her environment, and her instincts, desires and aversions tell her what to do in response to what she finds there. In fact, I believe that for the other animals, perceptual representation and desire and aversion are not strictly separate. Either through original instinct or as a result of learning, a nonhuman animal represents the world to herself as a world that is, as we might put it, preconceptualized and already normatively or practically interpreted. The animal finds herself in a world that consists of things that are directly perceived as food or prey, as danger or predator, as potential mate, as child: that is to say, as things to-be-eaten, to-be-avoided, to-be-mated-with, to-be-cared-for, and so on. To put it a bit dramatically – or anyway, philosophically – an animal's world is teleologically organized: the objects in it are marked out as being "for" certain things or as calling for certain responses. […] So these normatively or practically loaded teleological perceptions serve as the grounds of the animal's actions – where the ground of an action is a representation that causes the animal to do what she does. *(Korsgaard, 2011, pp. 10-11)*

We see that the familiar kind of normativity is already present in level 1 rationality. Here we do not have a clear distinction between different mental states, such as beliefs and desires; rather the worlds seems to be given to creatures as more directly organized in affordances. In other words, we might construe affordances as providing basic normative categories that are given to us in relation to our needs, preferences, and the tasks we are performing. Organisms that are susceptible to being evaluated in terms of level 1 rationality have much more flexibility in behaving, responding to environmental cues, and selecting action. Furthermore, the perception of affordances subserves the more fine-grained possibilities of classical and instrumental conditioning; that is, affordances provide the opportunity to affectively target specific actions in relation to specific circumstances of action. This enables organisms to more flexibly learn and adapt to changing environments, and to avoid the constraints of hardwired behavioral dispositions.

In accordance with this view, Bermúdez points out that level 1 rationality can also be evaluated in terms of short- and long-term criteria. Again, long-term criteria refer to maximizing inclusive fitness and short-term criteria refer to proximal goals that in the long run should support long-term goals. However, since at this level the flexibility of behavior and

plasticity in learning action-potentials is much greater, we can more often evaluate particular actions in relation to particular proximal goals. This flexibility that comes with level 1 rationality can account for the possibility of different criteria of rational evaluation coming into conflict.

For example, Bermúdez (2003, p. 121) points out that vervet monkeys have a complex system of signaling that enables them to warn each other when a predator is coming. Having such a system of signaling provides, in the long run, fitness benefits to every member of the vervet monkey population, as long as enough of them participate in the warning process.[100] However, participating in such a community provides the opportunity to behave in accordance with different rationality criteria. For instance, a vervet monkey, who, when faced with a predator, decides to flee, rather than staying and warning the others, might be acting rationally in more proximal terms, but not so rationally in terms of long-term inclusive fitness (as long as enough other monkeys play their part in the community).

As mentioned, at the top of the hierarchy we find level 2 rationality. The biggest difference at this level of organization is that the organism has the ability to flexibly respond to cues from the environment and the cognitive ability to step back, take into consideration its representations of the environment, and engage in a full-blown decision-making process. This is the level of cognitive sophistication in which a creature has the ability to become aware of the normal grounds of its actions and thoughts and therefore to take control over them (cf. Dennett, 2003, p. 204). When cognitive ability enables us to think reflectively, "[w]e are aware not only of our perceptions but also of the way in which they tend to operate on us" (Korsgaard, 2011, p. 11). Korsgaard furthermore writes that:

> once we are aware that we are inclined to believe or to act in a certain way on the ground of a certain representation, we find ourselves faced with a decision, namely, whether we should do that – we should believe or act in the way that the representation calls for or not. *(Korsgaard, 2011, p. 11)*

[100] These benefits are frequency-dependent because if most of the population does not warn other members when a predator is approaching, then it does not pay off to be the agent who warns others about danger and potentially risks her own life. However, if a great majority of the population participate in the warning process, then it becomes beneficial for some of the members to play the cheating strategy. In that case, non-reciprocators or cheaters get protection from others who make warning calls, but avoid the dangers of being injured or killed by providing warning calls themselves.

According to Korsgaard, this is the source of reason; the ability to reflectively think about the grounds, reasons, or 'rationales', as Dennett (2003) would call them, for our actions. A naturalistically conceived hierarchy of cognitive abilities suggests that for us a familiar kind of normativity already comes pre-packaged in our perceptions of affordances, and is not necessarily created at the level of self-reflective conscious reasoning. Nevertheless, for Korsgaard and others working in the Kantian tradition, it seems that reasons are individuated only in level 2 rationality.

According to Korsgaard, we take a consideration to be a reason "when we can endorse the operation of a ground of belief or action on us *as a* ground" (Korsgaard, 2011, p. 11). If we read this as stating that a necessary condition for something to be taken as a reason is for us to endorse it by representing it as a ground for our beliefs or actions then this would exclude level 1 rationality and affordances as providing reasons. The reason for this is that, according to Bermúdez, level 1 rationality does not presuppose decision-making that involves higher-order thought. There are at least two reasons for thinking that Korsgaard's view might not be right; one is empirical and the other more conceptual.

First, from a conceptual perspective, Korsgaard's view might lead to an infinite regress. As Peter Railton (2004; 2009) has argued in a similar context, if we presuppose that some consideration becomes a reason when we *endorse* it as a ground for action, the question is then what *endorsement* means in this situation. One natural proposal is to read it as some sort of action, possibly a (mental) *approval* on our part. However, read in this way, we naturally come to wonder whether this action of approval is legitimate or supported by reasons. If it is not, then we fail to see how that endorsement could make some consideration into a reason. But then, if supportive reasons are really normative reasons, they should be endorsed too, since rational endorsements make considerations into normative reasons. Since the question could be raised again at this point, we see how the infinite regress might be launched.

Alternatively, we could construe endorsement not as an action but as a sort of susceptibility or feeling that certain grounds count in favor of and that lead to some response (cf. Railton, 2004, pp. 194-195). However, if we grant this second reading, then we are in the ball park of level 1 rationality. As mentioned, counting in favor of, at this basic level, seems to be nicely interfaced with perceiving affordances. Or in this particular example, it is interfaced with having affective or otherwise intuitive responses that do not necessarily depend on our ability to self-reflectively think about the grounds of our thoughts and actions.

The idea that basic reasons come from the level 1 rationality also fits nicely with a naturalistic perspective. From an evolutionary perspective, agents with more complex decision-

making systems will be those that perceive affordances and have the ability to do something more. However, since the supposition is that agents of different complexities are on a motivational, affective, and cognitive continuum these more basic normative categories and perceptions of the world would retain their influence on the decision-making processes of more sophisticated reasoners.

This point can be illustrated by pointing to the phenomenon of moral dumbfounding (Haidt, 2001). For instance, when an average person is challenged to justify her judgment that incest is wrong, she usually searches for reasons that relate to harmful consequences that people who engage in incestual relations would suffer. Nevertheless, even when a psychologist who plays devil's advocate refutes all the reasons that pertain to show that incest is wrong,[101] people are still left with an intuition to the effect that incest is wrong. Jonathan Haidt (2001) says that when this happens people are dumbfounded – they have a strong feeling that there is something wrong with incest but cannot provide reasons for their judgments. The explanation for this is that for us the world is presented as already normatively circumscribed. Further down the cognitive line, these intuitions can feed into our more reflective deliberative system, where they compete with other intuitions and/or are evaluated in accordance with our other intuitions or criteria of reasoning that we adopt.

Nevertheless, Korsgaard and other Kantians are right in pointing out that what makes human agents different is their capacity for decision-making that, according to Bermúdez (2003), underlies level 2 rationality. Full-blown decision-making brings about different criteria for evaluating rationality. At the most general level, we find familiar criteria for judging instrumental or procedural rationality. This includes acting on the basis of reasons or grounds that are explicitly represented, such as when we act on the basis of an evaluation of the different consequences to which possible courses of action could lead. This includes assigning desirability values to these possible action-consequences and having instrumental beliefs about the likelihood of accomplishing different goals in accordance with their values. Decision-making can include choosing according to different criteria, not just those that depend on the consequences of a particular action. For example, deontologists (Gaus, 2011) point out that we can choose actions in accordance with the principles that we adopt; for instance, acting on an intention that can be properly universalized, or that is acceptable to all parties that are involved in a decision-making process, etc.

[101] For example, they might defend the couple engaging in sexual activity by saying that the intercourse would happen only once, partners would wear protection, everything is consensual, they love each other, etc.

The decision-making, and its elements, involved in level 2 rationality are susceptible to noticeably different criteria from levels 0 and 1 rationality. The ability of an organism to form detached representations of its environment and its value enables a more internally based evaluation of rationality. Here again, we can distinguish between more distal and proximal criteria of rationality. Distal criteria relate to fitness considerations, while the proximal become even more nuanced. For example, now we can evaluate particular mental states and their contents, regardless of how they correspond to reality, which introduces a higher possibility of conflicting judgments about the rationality of an agent. This explains the familiar phenomena that a person can be rational in her beliefs and actions, even though the action or the belief does not satisfy some externally given criteria (such as corresponding to reality, fitness benefits, actually satisfying an intended goal, etc.). For example, Mary may be rational in drinking from a glass full of petroleum, even if that would not be something that she desires or that would fulfill her other aims. The reason why she might be rational in drinking from the glass is because she believes that the glass contains gin and tonic (cf. Williams, 1981).

The possibility of conflicting criteria enables us to distinguish between reasons that come to us as normatively given, because of their individuation at level 1 rationality, and those that come from more sophisticated decision-making processes that involve more detached representations and evaluations of the environment. We can conceive of the relation between the two levels in the following way: the basic affordances that we perceive as external, along with other internally based instincts, will in a first step constrain our decision-making processes at a more cognitive level. What we see at the phenomenological level as *counting in favor of* will determine the values that we will try to pursue at a more cognitive level of decision-making. Thus, as a first approximation we might say that level 2 rationality will be evaluated in terms of how good it is at satisfying the goals set at level 1 rationality. Of course, our ability to contemplate our representations and what they stand for, and to take control over their grounds, will enable us to change the evaluations that come from a more primitive level.

To illustrate the point, we can think about how many people have implicit biases against people from other races. Nevertheless, we can suppress these biases, and even eliminate them through the top-down influence of our more cognitively sophisticated decision-making processes (see e.g. Kennett & Fine, 2009). However, the point that I want to make is that rather than thinking that top-down processes control everything, and that level 2 rationality criteria should dominate all others, we should be thinking about an interactive loop between levels. The idea is that primitive normative representations come from more primitive decision-making

processes and needs,[102] which then feed into the more cognitively based representational system, which *via* a feedback loop can influence these more primitive processes.[103] In this sense, the faculty of reason would be the whole thing that encompasses, on the one hand, more evolutionary and cognitively basic processes, and on the other, more cognitively and reflectively sophisticated ones. In this picture, substantive reasons are considerations that come from different levels of decision-making to mesh, compete, and provide grounds for more reflectively laden decision-making processes.

The important thing to note is that all three levels of rationality and the reasons for action they determine are fixed by external criteria. In other words, the criteria of rationality are fixed in reference to the task that we suppose the organism is performing and the abilities that it has (or that we can suppose that it has) in order to perform it. At levels 0 and 1, tasks are given by promoting fitness and other more proximal goals, such as feeding, mating efforts, avoiding predators, etc. While at level 2, we find many and possibly infinite varieties of tasks, since the human capacity for cognition enables us to think about abstract topics such as mathematical theorems, which do not necessarily relate to anything that is relevant for tasks related to maximizing fitness. Therefore, if we concentrate on level 2 rationality alone, then we have an indeterminate number of tasks that might provide a framework for judging rational action and thinking.

However, it could be objected that what has been said so far only applies to what we would consider to be motivational reasons, or at most reasons that are based on subjectively given ends and not on anything that we recognize as considerations that transcend individual-level authority, as supposed categorical reasons do. To answer this objection, in the next section I turn to considerations that will enable us to extend what has been said so far in order to accommodate the phenomenology of categorical reasons.

---

[102] For a discussion of the notion of need see Copp (1995, chapter 9). However, unlike Copp, I do not regard the introduction of needs in the account of reasons as being incompatible with a subject-based theory of reasons.

[103] This interactive feedback view is consistent with our contemporary understanding of the hierarchy of brain areas. For example, evolutionary more primitive areas underlying subcortical regions account for basic motivation and quick and automatic emotional responses. They provide inputs to the cortical regions above them, especially the prefrontal lobes, which evolved more recently and underlie higher-order cognition. The cortical regions then respond to impulses and regulate lower-brain areas, thereby composing a loop between higher and lower level brain regions (see e.g. Ardila, 2008).

5.4 Reasons and rational requirements

In order to provide more substance to level 2 rationality we need to think about the criteria or requirements that this type of rationality entails. As a plausible set of rational requirements that determine what reasons we have, Michael Smith proposes the following (where RR = reason requires that):

> R_1: RR (If someone has an intrinsic desire that p and a belief that he can bring about p by bringing about q, then he has an instrumental desire that he brings about q)
>
> R_2: RR (If someone has an intrinsic desire that p, and an intrinsic desire that q, and an intrinsic desire that r, and if the objects of desires that p and q and r cannot be distinguished from each other and from the object of the desire that s without making an arbitrary distinction, then she has an intrinsic desire that s)
>
> R_3: RR (If someone has an intrinsic desire that p, then either p itself is suitably universal, or satisfying the desire that p is consistent with satisfying desires whose contents are themselves suitably universal)
>
> R_4: ∃p∃q RR (If someone believes that p, then she has an intrinsic desire that q)
>
> R_5: ∃p RR (Rational agents do not desire that p)
>
> R_6: ∃q RR (Every rational agent desires that q) *(Smith, 2009, pp. 119-120)*

These requirements of reason are presented as being of increasing strength, starting from the weakest, R_1, to the strongest, R_6. R_1 and R_2 seem to account for reasons that we think are hypothetical, since these principles do not put substantive constraints on what our desires should be. R_1 is a familiar norm of instrumental or means-end rationality, according to which our goals set what we have a reason to do.[104] R_2 is a principle that tells us not to make decisions or form desires on the basis of arbitrary features of our goals. R_3 is a familiar Kantian principle that imposes a universalization constraint on what type of motivations or intentions we can act upon; and could be seen as an intermediate principle between the purely hypothetical and strictly categorical ones. R_4, R_5, and R_6 could be seen as most clearly falling under categorical reasons, since they demand that rational agents have particular desires and consequently that they be

[104] The norm of instrumental rationality is usually construed as being a part of procedural rationality more broadly construed, where procedural rationality also includes principles for correct and reliable belief-formation, such as different forms of deductive and inductive inferences, probability theory, etc. (Bermúdez, 2003, pp. 110-111; Smith, 2012, p. 234).

disposed to perform certain actions no matter what motivational set they have to begin with. An example of R_4 could involve forming desires and intentions on the basis of normative beliefs, for example believing that it is wrong to hurt other people gives you a reason to desire not to hurt other people and to form your intentions in accordance with that norm. Parfit (2011a) forcefully argues for something like principles R_5 and R_6 when he claims that the intrinsic nature of future agony provides one with a reason to desire to avoid it. Other examples could include the more common idea that when people are harmed or injured then other people have reasons help them if they are in a position to help.

Unfortunately, the issue of whether the presented principles are in some sense valid is controversial (Smith, 2009, p. 124). Some authors think that principles of rationality are minimal, resembling R_1, while others think that rationality can be very substantive, admitting principles as strong as R_6. I believe that part of the controversy lies in the fact that many authors think that if these requirements of reason are valid they need to be justified by *a priori* considerations.

For instance, Smith (2012, pp. 238-239) contends that if something like R_1–R_6 provide principles of rationality then we should be able to derive them through *a priori* reasoning. Since many authors have doubts about the possibility of showing *a priori* that there are desires that everybody should have regardless of their starting points (Railton, 1986; Williams, 1981; 1995) it is argued that only principles of the form of R_1 could be unproblematically granted an *a priori* status (see e.g. Callebut, 2007, p. 80). However, from a naturalistic point of view, even the *a priori* validity of the instrumental requirement could be challenged.

This possibility might seem strange because the plausibility of having a desire to p seems to be conceptually connected to being disposed to take the means you believe to be necessary to accomplish p. Nevertheless, this conceptual construal of what it is to have a desire needs to be distinguished from the proposed principle of rationality R_1. According to Smith, for something to be a principle of rationality it needs to tell us "how to reason when we deliberate" (Smith, 2009, p. 121). If R_1 or its variants are norms that one should be able to follow in reasoning about what to do, then it is possible that there are environments in which reasoning in accordance with R_1 will not lead one to accomplish one's goals or tasks.

To illustrate this, consider an example adopted from Morton (2010, p. 569; see also Broome, 2007, pp. 173-174). Imagine a world in which there is an evil demon whose aim is to make your life difficult. In fact, whenever you deliberate and form beliefs about the necessary and sufficient means to accomplish your ends, the demon changes the circumstances in the world so that your beliefs do not lead you to successfully perform actions by which you could

satisfy your ends. Let us also suppose that you are so attuned to your environment that following your instinct will most often lead you to successful action. We could suppose that your perception of affordances is so good that you can in most circumstances act successfully without deliberating about what the necessary means or what desires or intentions to form. In such a world, it would not benefit you, and thus you would not be justified in following the norms of instrumental rationality. Rather, following your instincts would be a better strategy most of the time.

This example illustrates that there is *prima facie* difficulty in accounting for the *a priori* status of the instrumental norms of rationality. Since we face this difficulty even for the basic norm that involves means-end reasoning, we can also be skeptical about the prospects for offering *a priori* justification for other more substantive norms of rationality. From a naturalistic perspective this is not surprising. According to this perspective, what we think we have a reason to do and the validity of those beliefs depends on our contingent natures and the environments we are faced with; for example, it depends on experiences, learning histories, cultural background, reasoning abilities, etc. Furthermore, the notion of a rational person should be interpreted as being humanly rational. This, in the light of the previous discussion on levels of rationality, should be further specified in relation to the task that humans have phylogenetically or ontogenetically evolved to perform and the environmental and cultural niche to which they are adapted.

To account for the possibility of categorical reasons, rather than trying to show how particular norms became categorical reasons for people to act upon, I will provide a model that will pertain to show how this phenomenon could have arisen in general, without relying on *a priori* intuitions about what particular reasons we actually have.

5.5 The emergence of categorical reasons

In order to show how categorical reasons could be naturalistically accommodated, I will start by examining in more detail how reason relations plausibly come to be formed in the first place. The upshot of this discussion should be to show that hypothetical and categorical reasons are not distinct in kind, but rather lie on a continuum of reasons that are more or less dependent on particular individuals' preferences, beliefs, values, etc.

It is important to note that already at the level of affordances, things that are presented to us as counting in favor of something will often not be phenomenologically construed as depending on us or being in a broad sense subject-based. When I realize that my life is in danger,

I do not see this situation as demanding some response from me because I perceive myself as being a person who has a standing desire or a goal to avoid danger. Rather, we would generally see the situation as being such that it demands some response from us or counts in favor of our avoiding danger. Paradoxically, perhaps, situations lose this sort of primitive normativity when we get to the level of reflective rationality and start thinking dispassionately about things, such as when I start to ask myself whether I should avoid danger, whether I should be the sort of person who always plays it safe or whether I should take more chances in life, etc.

Thus, even at this basic level reasons are not presented to us as being based on our subjective needs. However, the question remains: when we get to a more reflective level some situation's normativity would seem to depend on our having certain desires and goals, while others would seem normative regardless of our particular aims. I think that this distinction between reasons can be explained in the same way in which naturalistically minded authors explain the formation of semantic relations more generally. The basic idea is that the establishment of certain primitive semantic relations is homomorphic or even isomorphic to establishing certain reason-relations.

5.5.1 Primitive semantic content and reasons

William Harms (2004), following Millikan's (1989) teleological semantic program, develops a naturalistic framework in which he explains the emergence of basic semantic features of indicative and imperative or normative contents of different semantic units, and the origins of basic normative intuitions about how things should function in general. I propose applying this framework to normative reasons.

In this framework, the basic concept is that of a primitive content. Primitive content involves representations that have a double function: they function to *indicate* that things are such and such, and at the same time they pertain to show which actions *should* be performed. The idea that certain representations have primitive contents is similar to representations that Millikan (1996) calls 'Pushmi-Pullyu Representations.' Paradigmatic examples of this kind of representations include vervet monkey warning calls or the characteristic dance of honeybees. For instance, the vervet monkey's warning call has at the same time an indicative function, by which it indicates that a predator is approaching, and a directive function, signaling that other monkeys should run away. Similarly, the honeybee's characteristic waggle dance has the functions of indicating where the foraging or habitat resources are and of telling other bees how far away they are and in which direction they should fly.

The important thing to note is that the basic meaning of biological signals, like language, is established by convention, that is, by conventions that determine in which situation it is appropriate to produce a signal and conventions that determine which action or response is appropriate as a consequence of the signal in that situation. For example, two aspects of the meaning of the term 'water' are its reference and implications. For instance, 'water' refers to the H_2O molecule, but it implies that the thing to which it refers is relatively transparent, quenches thirst, can be used for washing, etc. These two aspects of representations Harms (2004, p. 193), following a venerable philosophical tradition, albeit in a more circumscribed way, calls extension and intension. The extension of a representation is what the representation stands for or is representation of, such as a thing or a possible state of affairs. Intensions are what follow from a proper use of representations, which is determined by their roles and relationships to other representations in a representational system.

> In human language, these include the definitions of terms (which are often taken to determine their extensions), the logical implications of sentences, the 'modes of presentation' (like attributing beliefs rather than expressing them), and various attitudes one can have toward propositions (e.g., believing that p, hoping that p), which together weave the collection of signs and symbols into a representational system. *(Harms, 2004, p. 194)*

As already indicated, representations do not have to be expressed in a linguistic form. So according to this picture, basic signals, such as warning cries and bee dances have meaning, and therefore extension and intension in the present sense. According to Harms, the conjunction of a representation's extension and intension constitute its *content*. In primitive contents, representations have both indicative and directive functions. In more sophisticated representations such as beliefs and desires these two functions can come apart, since belief's extension and intension will most often have an indicative function, while desire's extension and intension will play a purely directive function.

For our present purposes, it is important to note the features that make representations and their contents analogous to reasons or facts that count in favor of something. First, representations have extensions, which are standardly taken to be truth-conditions. Reasons have grounds, that is, facts, state of affairs, or true propositions that form the grounds of reason-relations. Second, representations have intensions, that which follows from the role that they play in a representational system in relation to the conditions that form their extension. Similarly, reasons are reasons for something, whether an action or an attitude. Third, reasons seem to have a double function too. They indicate what seems to be the case, but at the same time indicate what should be done in response to the situation. Thus, reasons seem to have

similar features to primitive semantic content. To remind ourselves, primitive semantic content at the same time plays an indicative and directive role. Reasons seem to be constructed of the same two things: they have grounds, so they indicate how things are, but also on the basis of these grounds they are directed towards some response or reaction.[105] Thus, for our present purposes I propose to identify the reason-relation with representations or a subclass of representations that have the phenomenology of *counting in favor of* (for a similar suggesstion see also Harms & Skyrms, 2008, pp. 444-446).

The analogy between reasons and primitive semantic contents will enable us to see how categorical reasons can be based on naturalistic ingredients. To start with, the establishment of basic semantic relations between signals and responses can explain how familiar hypothetical reasons, that is, those that depend on the goals of an agent, emerge. How meaning conventions get established is standardly explained in terms of a game-theoretical model, as influentially proposed by David Lewis (1969) and then, most notably, further developed by Brian Skyrms (1996; 2010). A simple model can describe the establishment of meaning conventions or how a signal acquires particular meaning.

We start by examining a cooperative game with two players or agents.[106] Those agents can play two roles in the game: one can be a sender (S), who sends a signal, or a receiver (R), who receives a signal and thereby responds to it by acting in a certain way. The roles are not prefixed, so some of the time an agent will play the (S) role and on other occasions the (R) role and *vice versa*. In the basic construction of the game, the agents have the possibility of perceiving two states of the world (W_1 and W_2), they can send two messages (M_1 and M_2), and react by performing two different actions (A_1 and A_2). Performance of A_1 is correct for circumstances W_1, and A_2 is correct for W_2. The basic idea is that if one player correctly responds to a message sent by the other player in response to detecting some state of affairs, then both of them receive a positive payoff $a>0$ (i.e. the payoff is a number that is greater than 0), otherwise they get

[105] We can note a further factor that also supports the analogy. Representations can compete for a response, just as conflicting reasons, depending on their weight, can compete for a response. For example, in a Stroop task people are presented with color words that are differently colored. The task is to say, in a short time span, the color of the word. If the word is red, for example, but it says green, people tend to be biased towards saying that the word is green even though it is actually red. The reason for this seems to be the competing representations that people have of the same situation. Psychopaths, on the other hand, do not show this task bias, and perform better than non-psychopaths, because it seems that their perceptual system does not experience conflict between competing representations (Zeier, Maxwell, & Newman, 2009).

[106] The following exposition and notation relies on (Harms, 2004, pp. 194-195) and (Huttegger, 2007).

nothing (the payoff is 0).[107] The supposition is that every action is a correct response to only one state of affairs. Thus, to respond correctly is to ensure a coordination between a single state of affairs and a single action. The game is set such that only the sender perceives the state of affairs and sends the signal to the receiver, and the goal of the game is accomplished if the receiver responds in a way that is appropriate given the situation in which the signal is sent (i.e. if the payoff is some number greater than 0).

Since, in the basic case, the sender and receiver do not have a preestablished system of communication, there are four sender and corresponding receiver strategies that players can execute. These are given in **Figure 3** .

Sender strategies	**Receiver strategies**
S_1: M_1 if W_1; M_2 if W_2	R_1: A_1 if M_1; A_2 if M_2
S_2: M_2 if W_1; M_1 if W_2	R_2: A_2 if M_1; A_1 if M_2
S_3: M_1 if W_1 or W_2	R_3: A_1 if M_1 or M_2
S_4: M_2 if W_1 or W_2	R_4: A_2 if M_1 or M_2

Figure 3 *(adapted from Harms, 2004; Huttegger, 2007)*

Given the role of the agent, she can combine strategies so that, for instance, when she is sender she can execute S_1 and when she is receiver, she can execute R_1. Thus, every agent can combine sender and receiver strategies, depending on the role she plays. For instance, she can combine S_1 with R_2, S_2 with R_1, S_2 with R_2, etc. (there are 16 possible combinations of strategies). In this example, our interest lies only in two combinations of strategies, S_1R_1 and S_2R_2 (see **Figure 4**), since they bring maximal payoff to the agents (Harms, 2004, pp. 195-196). In technical terms, they constitute a Nash equilibrium. That is, when either of these two strategies is established, no one agent has a unilateral incentive to stop playing them. These combinations of strategies manage to do this because they put states of affairs, messages, and actions into one-to-one relation. In this sense, if both agents coordinate on one of these two combinations, they will always benefit from their interactions, that is, their responses will be the best they can be in relation to what the other agent is doing.

[107] Depending on the type of interaction that the game is modeling, payoffs can be construed as desire satisfaction, fitness benefits, or something similar.

We see how the meaning of the message is conventional in the signal system that gets established. If players settle on S_1R_1, then M_1 would indicate that the world is in state W_1 and that A_1 should be performed. If they settle on S_2R_2, then M_1 would mean that W_2 is the case and that A_2 should be performed. But more importantly for the present context, this example illustrates how reasons could emerge from interactions between agents. When the meaning-convention is established, then, in this simple case, we can say that a reason-relation is established as well. For instance, if S_1R_1 provides a signal system for when to perform actions A_1 and A_2 then we can say that being in W_1 gives a reason or counts in favor of performing A_1.

Reason requires (RR)	
S_1R_1	**S_2R_2**
$W_1 \rightarrow RR \rightarrow M_1 \rightarrow RR \rightarrow A_1$	$W_1 \rightarrow RR \rightarrow M2 \rightarrow RR \rightarrow A_1$
$W_2 \rightarrow RR \rightarrow M_1 \rightarrow RR \rightarrow A_2$	$W_1 \rightarrow RR \rightarrow M_1 \rightarrow RR \rightarrow A_2$

Figure 4 *(adapted from Harms, 2004, p. 196)*

The model is naturally construed as applying to an interaction between different agents. However, there is nothing formal that prevents the application of the model to single agents, in the sense that we can explain how particular representations in a single system acquire their meaning or how single reason-relations for particular agents get established. For example, S_1 can be implemented by an agent's perceptual system, and R_1 as a system that produces actions in response to signals coming from S_1. Similarly, when the perceptual system produces signal M_1, an agent will see this as a reason or something that counts in favor of performing A_1, whether that is an action or some other belief (depending on our interpretation of elements of S_1R_2).

To return to the interpersonal case, we can see how categorical reasons can emerge from simple associations between efforts to coordinate actions. Once enough of the population play the strategy S_1R_1, for instance, it will become rational for every other agent who is inclined towards cooperation to regard W_1 as a reason to A_1. That will be the case no matter what the occurrent preferences or beliefs of that agent might be. Seen from an evolutionary perspective, the cooperative efforts of many generations of agents will produce a system of reason-relations that new agents will simply grow into, and many of those reason-relations will simply be experienced as things that count in favor of producing appropriate responses. They will be experienced as such without providing explicit or transparent explanations for why this is the

case (for this we would need to examine the history and the evolution of the individual or the society of agents). Such as when we see a person in pain we understand that she has been hurt and that this situation *demands* that we respond by helping her in some way. However, the explanation of why this particular fact counts in favor of performing this act will be different depending on the normative narrative that different people accept about the origins or groundings of this relation.

As mentioned, categorical reasons will emerge from interactions between agents similarly to the way in which reasons emerge at the level of a single agent, by establishing associations between state of affairs and the responses that bring some benefit in relation to those state of affairs. However, this will happen only if enough other agents behave in similar ways and obey similar associations between states of affairs and actions. In this sense, interpersonal categorical reasons are frequency-dependent. They will emerge and be stabilized only if at the level of a population of agents enough of them act cooperatively and at least in the long-term benefit from the cooperation.

So how does this picture explain the difference between hypothetical and categorical reasons? My suggestion is that when we come to think reflectively, whatever reasons are provided by our personal goals and desires, we become prone to seeing them as optional and not really externally binding. This may be because personal reasons depend on our contingent plans and desires that are often ephemeral or the products of different quirks, which we can, by exercising self-control, influence, change, and/or come to deem invaluable. However, when we think about social norms, especially those that relate to our well-being and the well-being of others, we do not see them as optional because we cannot influence them just by exercising self-control, for example. In fact, we see them as providing a platform, in accordance with which we control our behavior. This non-optionality comes from the fact that they are given to us as external to our particular motivational sets. Nevertheless, when we observe the situation from an evolutionary point of view or with respect to the way in which reason-relations emerge we see that there is no qualitative difference between hypothetical and categorical reasons. Categorical reasons at the level of an individual can be viewed as hypothetical at the level of a population of agents whose strategies are in a stable equilibrium. This is not because some other state of affairs – W_2 instead of W_1 – could have been a reason for performing A_1, it is because the nature of the agents and the nature of the interactions between them that makes certain states of affairs into categorical reasons for doing something.

5.5.2 The role of rationality and normative intuitions

From this perspective, we can explain the role of rationality and normative intuitions about what counts in favor of what. Harms (2004, p. 206) supposes that normative intuitions are the outputs of higher-order cognitive or affective systems that take as inputs violations of the functions of lower-level systems and output a response that reinforces the lower-level rule. In our case, on the basic level we have established reason-relations, the mechanisms that process them, and the higher-level systems that regulate and reinforce functions of the former. For instance, one of the most basic and general requirements for successful cooperation is to obey the norm of reciprocity. If you do someone a favor you expect to get something in return, especially if the favor is significant. Regulation of these interactions is underlined by our intuitions regarding what is fair (Baumard, André, & Sperber, 2013). So, most agents who are disposed to cooperate will feel that there is a basic reason, when you do a significant favor for someone, to expect something in return and *vice versa*. When you notice that somebody is trying to cheat, for example, by receiving a favor but not giving anything in return, then the intuitions underlying fairness signal that there has been a breach of the norm, that the reason was not obeyed, so to speak. These intuitions signal that punishing behavior is appropriate; for instance, you warn the cheater, report him/her to the relevant authorities, etc. In other words, the intuitions reinforce the basic mechanism that processes and satisfies the reason-relation. The same thing could happen when we cheat in some way and our conscience starts to bother us. This can also be viewed as a reinforcing intuition, the only difference being that in this case the punishing signal is directed towards oneself.

Similarly, epistemic intuitions regulate how we should reason and form beliefs (Harms, 2004, p. 206). This is most noticeable when we find ourselves confronted with two inconsistent beliefs. The intuition that coherence is violated forces us to abandon one of the beliefs. Usually, the one that gets dropped is the one that is less entrenched in our belief or knowledge database. Epistemic norms also have an important social function (Mercier & Sperber, 2011; Smokrović, 2015). Communication can potentially be beneficial, but an agent needs to be able to decide whether a piece of information is credible or not. Rather than relying on intuitions alone, an agent needs to be able to evaluate arguments given by others and to produce plausible arguments that will be convincing to other people. This involves employing overtly rational capacities in order to properly evaluate and respond to the evidence at hand.

Thus, we see that rationality, in the sense of level 2 rationality, also enables us to respond to reason-relations that are already established. It enables us to respond to them in a more

flexible way than automated intuitions. For example, it enables us to more effectively protect ourselves from possible cheaters and to more effectively enforce the rules of fairness. In addition, the possibility of detachment from our present motivations and representations that goes along with reflective rationality enables us to evaluate the current reason-relations that we adopt and to see whether some better normative relations could be established in the light of other reasons that we endorse. I will close this chapter by illustrating this last point.

Mindless evolutionary processes can to lead to many different equilibrium points and therefore can establish many different reason-relations. Consider a modified signaling game in which there is a partial conflict of interest between senders and receivers (Zollman, Bergstrom, & Huttegger, 2013).[108] In this situation, senders can be of two types, T_1 and T_2, and they either can send a signal or not send a signal. The receiver has two possible actions, A_1 and A_2, that are correct responses to signals coming from types T_1 and T_2, respectively. The receiver cannot determine which type of sender she is playing. So in choosing the appropriate action she must rely on whether the signal is sent or not. There are four sender strategies and four receiver strategies available (see **Figure 5**). In this game, there is partial conflict of interest, because if the sender is of type T_1 and sends a signal, then both the sender and the receiver will benefit if the receiver performs A_1. However, if the sender is of type T_2, then it will still benefit her if the receiver, by reacting to a signal, performs A_1 – though this would not benefit the receiver, because the right action to perform in response to T_2 signals is action A_2.

Sender strategies	**Receiver strategies**
S_1: signal if T_1; do not signal if W_2	R_1: A_1 if signal; A_2 if no signal
S_2: do not signal if T_1; signal if T_2	R_2: A_2 if signal; A_1 if no signal
S_3: always signal	R_2: always A_2
S_4; never signal	R_4: always A_1

Figure 5 *(adapted from Zollman, Bergstrom, & Huttegger, 2013)*

To illustrate the functioning of the game, we can imagine that senders are people who ask for social benefits and that they differ by their social and economic status. Type T_1 are those who belong to a lower socio-economic group and T_2 are those who belong to a higher socio-

[108] In what follows, I will describe a version of the so-called Sir Philip Sydney game, developed and used by John Maynard-Smith for modeling evolutionary interactions between animals that have partially different fitness-interests (Zollman, Bergstrom, & Huttegger, 2013).

economic group. Receivers could represent institutions whose job is to appropriately and justly (since resources are limited) grant financial and other types of help to people from the appropriate group. Thus, receivers either grant requests (A_1) to people of type T_1 or refuse to grant help (A_2) to people of type T_2. Nevertheless, since there is no cost in sending a signal no matter what type of person you are, it is still beneficial for T_2 people to send signals and reap the ensuing benefits, which stem from the inability of receivers to discriminate between types of people without relying on signaling cues.

In the situation where there are no signaling costs, it seems that even by spontaneous evolution most people, when in the sender role, will tend to play the S_3 strategy. When in the receiver role, they will probably tend to play a combination of R_1 and R_2, since by only playing R_1 resources would be soon depleted. Let us suppose that, in response to S_3, receivers come to play strategy R_1 60% and R_2 40% of the time. In fact, if there are no signaling costs, the reason-relations that would emerge would be of a certain strength, since 60% of the time signaling would count in favor of doing A_1 and rest of the time it would count in favor of doing A_2. And everybody who joined the game would tend to react to these reasons appropriately.

Now, let us suppose that receivers and senders develop rational capacities that enable them to detach from their current representations and motivations and think about the present situation more globally. Receivers and senders of type T_1 would realize that there is a better equilibria of strategies in the vicinity, namely, those that include combinations S_1R_1 and S_2R_2, and they would start thinking about moving their interactions more closely to these equilibria. How they would achieve this move to a better equilibrium? First, receivers would start to be vigilant by creating costs for senders that deceive by signaling inappropriately. This could include not taking the signal at face value, investigating where the signal comes from; they could argue and ask for reasons or justifications from senders; those senders that are caught sending deceptive signals could be ostracized or punished by having their benefits taken away, and so on and so forth. Second, those belonging to type T_1, who are deprived of the benefits, would probably participate in denouncing cheaters and indicating that there is a better equilibria of interactions that is worth pursuing. Thus, in this way, deploying reason or rationality would abolish the validity of old reason-relations or indicate their falsity. Furthermore, using reason would help to indicate which norms to create or how to reach more stable and effective equilibrium points.

5.6 Summary

In this chapter, the goal was to further develop one type of subject-based theory of reasons. In particular, one of the main goals was to indicate how categorical reasons could emerge and how their existence could be explained in terms of a subject-based theory of reasons. In order to do this, I touched upon different topics, including the relation between reasons, the faculty of reason, and norms of rationality. I argued that from a naturalistic perspective it makes sense to try to explain reasons in terms of the faculty of reason and the principles that govern it. However, inside this framework I distinguished between different types and criteria of rationality and their relation to reasons. I indicated how, by using a model from game theory, categorical reason-relations could emerge. Finally, in this framework, I tried to explain how rationality construed as a reflective ability that enables agents to detach themselves from their occurrent motivations and representations could respond to reasons or even establish new reason-relations.

6 Rationality in practice: Psychopathy as a case study

6.1 Introduction

The role of this chapter is to apply some of the considerations relating to reasons and rationality discussed earlier in the thesis in a more practical setting. Recently, there has been interesting discussion about whether people with psychopathic personality disorder are, by the nature of their condition, comparatively more irrational than other people. The answer to this question is significant because it could have interesting philosophical and legal implications. The task of this chapter is to apply the framework developed earlier in this thesis in order to see how the debate on the rationality of psychopaths could be fruitfully framed and what conclusions we are currently justified in reaching about their rationality.

The chapter is organized in the following way. In the remainder of the introduction, I will briefly say why the question of the rationality of psychopaths is important for different philosophical and more practical debates. In section 6.2 I will explain in more detail how psychopathy is measured and look at the abnormalities usually correlated with it. I will then present an influential argument according to which empirical data show that psychopaths are instrumentally irrational. As we will see, a proper evaluation of the argument requires making some conceptual clarifications concerning the criteria of rationality that are being applied. In section 6.3 I will evaluate the argument by arguing that the relevant notion of rationality is internal rationality. In section 6.4 I will evaluate the argument in an alternative way by applying the notion of external rationality. The overall conclusion is that, based on the notions of rationality that we apply (and their ramifications), we are not currently justified in concluding that psychopaths are instrumentally irrational.

6.1.1 The significance of psychopathy for philosophy

Psychopathy is standardly characterized as a personality disorder involving severe affective and interpersonal deficits that reoccur across different cultures (Cooke, 1998). The issue surrounding psychopathy is most notably related to the antisocial behavior that seems to

accompany psychopathic personality traits. Psychopathic individuals are disproportionately more likely than any other group of people to commit a crime and violently recidivate (Kiehl & Hoffman, 2011). In this respect, they put enormous pressure on our moral, legal, and economic systems.

The issue of whether psychopaths are rational is important for numerous debates. For instance, sentimentalists about moral judgment have argued that psychopaths exemplify people who are rational but immoral (Aaltola, 2014; Nichols, 2004; Prinz, 2006). Moral rationalists, instead, dispute the claim that psychopaths are rational (Kennett, 2010; Maibom, 2005; 2010). Furthermore, psychopathy has been used as a case study in the debate about internalism and externalism about moral judgment; externalists have argued that psychopaths are rational and possess moral understanding, but are not motivated to act morally. Internalists have responded by disputing the claim that psychopaths are rational and make moral judgments (Sinnott-Armstrong, 2014).

However, in the more applied domain the rationality of psychopaths is central to the question of whether they should be held culpable or responsible for their wrongdoing. If a psychopathic offender satisfies the criteria for moral and criminal responsibility, then she may be deemed culpable for her wrongdoing and subject to appropriate punishment by our penal systems. If she were to be found to be unable to satisfy the criteria for being responsible as prescribed by our moral and penal systems then the appropriate social response to psychopathic offenders would perhaps include medical or another kind of treatment rather than penal punishment (Nadelhoffer & Sinnot-Armstrong, 2013, pp. 229-230).

Among the minimal criteria for moral and criminal responsibility, we find the idea that a person needs to be minimally rational (Sifferd & Hirstein, 2013). According to Stephen Morse, an influential philosopher of law, minimal rationality includes, on the one hand, having epistemic competence, that is, the ability to form, revise, and justify beliefs. On the other hand, it involves being able to reason instrumentally, where this includes "weighing the facts appropriately and according to a minimally coherent preference-ordering" (Morse, 2000). Minimal rationality sounds very much like the principle of reason that, following Smith (2009), I named R_1.[109] Thus, if it could be shown that psychopaths are not rational in this minimal sense, then this could potentially have great repercussions on how the law should treat them when they do something illegal. In fact, there are authors who argue that the rational capacities (in this minimal sense) of a typical incarcerated psychopath are impaired (Maibom, 2005). Based on

[109] The only difference is that in Morse's formulation the epistemic part is made explicit.

similar considerations, others have argued that psychopaths should therefore be liable to an insanity or diminished responsibility defense (Sifferd & Hirstein, 2013). Heidi Maibom (2005) provides an influential argument to the effect that psychopaths are irrational in this thin, instrumental sense, and in what follows I will discuss her argument in the light of the framework that I developed in the last chapter.

6.2 Measuring psychopathy: PCL-R

Psychopathic personality disorder is a condition that is significantly associated with criminal behavior. There are many different measures of psychopathy, however the most influential diagnostic tool that is widely used in scientific and forensic settings is the so-called Psychopathy Checklist, which was developed by Robert Hare. According to Hare's Psychopathy Checklist-Revised (2003), psychopathy is a personality disorder characterized by Glib/superficial charm, Lack of empathy, Grandiose sense of self-worth, Conning/manipulativeness, Lack of remorse or guilt, Parasitic lifestyle, Poor behavioral controls, Early behavioral problems, Lack of realistic long-terms goals, Impulsivity, Irresponsibility, etc. (see **Table 2**).

Table 2 PCL-R items

Factor 1	**Factor 2**
Interpersonal	**Lifestyle**
1. Glibness/Superficial charm	3. Need for stimulation
2. Grandiose sense of self-worth	9. Parasitic lifestyle
4. Pathological lying	13. Lack of realistic, long-term goals
5. Conning/Manipulative	14. Impulsivity
	15. Irresponsibility
Affective	**Antisocial**
6. Lack of remorse or guilt	10. Poor behavioral controls
7. Shallow affect	12. Early behavioral problems
8. Callous/Lack of empathy	18. Juvenile delinquency
16. Failure to accept responsibility	19. Revocation of conditional release
	20. Criminal versatility

The items on the list are evaluated by awarding points from 0–2, which means that the maximum number of points that one can get on PCL-R is 40. According to PCL-R as it is used in United States of America, a person that scores 30 or more points is psychopathic, while in some European countries the score for diagnosing someone with psychopathy is much lower, for example 25 or more points.

Empirical studies have shown that psychopaths exhibit impairments in empathy; for example, they have reduced autonomic responses to the sadness of other individuals. In recognition tasks, they also show impairments in recognizing sad and fearful facial expressions and vocal affect (Blair, Mitchell, & Blair, 2005, pp. 54-55). In addition, this body of evidence in favor of emotional impairments in psychopathic individuals has been corroborated by neuropsychological and neuroimaging studies. These have shown that the underlying causes of psychopathic disorder might stem from an abnormal functioning of the paralimibic system (Kiehl, 2008), the part of the brain that regulates processing of affective information. In these neuropsychological studies the amygdala and ventromedial prefrontal cortex, the brain regions that are involved in emotion regulation and representation of the affective value of stimuli, respectively, have been implicated as likely defects that underlie the psychopathic disorder (Blair R. J., 2008).

6.2.1 The rationality of psychopaths

However, emotional deficits are not the only differences that psychopaths exhibit. There are studies showing that psychopaths suffer from decision-making and cognitive deficits too. Using these data from neuropsychological tasks, some authors, most notably Maibom (2005; 2010; see also Sifferd & Hirstein, 2013), have argued that these decision-making deficits indicate deficits in psychopaths' practical rationality. In fact, Maibom (2005) has argued that psychopaths have a very basic deficit in instrumental rationality; according to her, they often fail to desire the necessary means to their desired ends.

Maibom (2005) grounds her conclusion on the basis of a plausible principle of instrumental rationality, on the one hand, and on decision-making studies in psychopaths, on the other. Following the Kantian tradition, Maibom introduces a list of rational requirements:

> Practical rationality requires that who wills the end, also wills:
>
> (a) the means that are indispensably necessary to his actions and that lie in his power,
>
> (b) some sufficient means to the end,
>
> (c) to make available necessary and/or sufficient means to the end if such means aren't already available,
>
> (d) that the various specific intentions that are involved in adopting a maxim are mutually consistent, and
>
> (e) that the foreseeable consequences of acting on the specific intentions are consistent with the underlying intention. (Maibom, 2005, p. 241)

Maibom argues that psychopaths are irrational because they have problems obeying all of these principles. I will focus only on the first principle, since if psychopaths really have a problem in satisfying (a), then it is plausible to expect that they will have problems in satisfying the rest – though the reverse does not necessarily hold.

Maibom (2005) argues that psychopaths' performance in so-called instrumental learning tasks can be taken to show that they do not satisfy (a). Instrumental learning tasks involve learning, in a certain situation, the relation between stimuli and the appropriate response through administration of rewards or punishments.[110] There are many versions of these tasks; I will only mention two as most illustrative for our present purposes. In response-reversal tasks subjects are asked to respond (e.g. to press a button) if they see one type of stimuli on a computer screen (e.g. a square) and to withhold from responding if they see some other type of stimuli (e.g. a circle). Furthermore, if they respond correctly they get a reward (e.g. money), but if they respond incorrectly they are punished (e.g. lose money). Once the relation between the stimuli and the expectation (of reward or punishment) is reinforced, the experimenter changes the reinforcement contingencies, so that the response that was previously rewarded, in the reversal condition is punished. In this type of task psychopaths often fail to change their responses, and continue to respond to the stimuli that was previously rewarded even though it is now punished (Blair, Mitchell, & Blair, 2005, pp. 100-101). Nevertheless, they do learn to respond correctly

[110] Rewards and punishments usually involve money that is either earned if the response is correct or lost if the response is not correct.

to changes in the reinforcement contingencies, although at a slower pace than control groups (Brazil, et al., 2013).

Psychopaths also perform worse than controls in the so-called Iowa-Gambling task (Blair, Colledge, & Mitchell, 2001; Mitchell, Colledge, Leonard, & Blair, 2002). This paradigm consists of a card game in which participants are asked to select cards from four decks (A, B, C, and D) that are presented on a computer screen. Each of these decks are associated with different monetary rewards and punishments. In order to earn money until the end of the game, participants need to learn to select cards from overall winning decks and learn to avoid disadvantageous decks. For instance, decks C and D have a higher frequency of punishment, but by the end of the game choosing from these decks will earn you money. Choosing from decks A and B is disadvantageous, since choosing only from these decks will leave you with less money than you started with. Thus, participants need to learn to choose from decks C and D. The trick is that at the beginning of the game, choosing from decks A and B provides you with sizable gains in money. However, continuing to select from A or B leads to even greater losses. Thus, a person needs to learn to avoid A and B and to turn to C and D, which do not enable big gains, but at the end of the game still allow her to earn some money. Psychopathic participants, compared to controls, show non-risk-aversive behavior in selecting cards from disadvantageous decks throughout the task. Accordingly, they sustain major losses.[111]

According to Maibom, response-reversal errors and failures in the gambling task indicate that psychopaths do not will the necessary means for their ends. The adduced reason is that in the first task on average they learn less well than controls to change their responses in relation to the changing environment, while in the second task they learn less well to choose from rewarding decks of cards. However, according to Maibom, learning in these cases presents the necessary means for accomplishing their ends (Maibom, 2005, pp. 242-243).

Even though instrumental learning tasks show something interesting about psychopaths' decision-making processes, they do not by themselves show that psychopaths are irrational in some substantive sense. In order to show that psychopaths' performances on instrumental learning tasks indicate something about their instrumental rationality, we need to be clear about what requirement (a) actually demands from rational agents and under which conditions.

[111] It seems that healthy people learn to advantageously choose from C and D because their emotional system negatively marks big losses stemming from the decision to select cards from decks A and B (Bechara A. , Damasio, Damasio, & Anderson, 1994). According to one influential hypothesis, since psychopaths do not display normal emotional processing of information their emotional marking system does not mark bad decks negatively, so they continue to be prone to risky behavior by choosing from bad decks of cards.

To remind ourselves, according to requirement (a) rationality requires that whoever wills the end also wills the means that are indispensably necessary for attaining this end and that *lie in his power*. Here it is important to notice the proviso "lie in his power." Maibom does not explicitly say what the proper interpretation of this phrase might be. Onora O'Neill, from whom Maibom adopted the list of rational requirements, interprets this phrase as being about *available* means (O'Neill, 2001, pp. 311-312). Given the discussion of levels of rationality in chapter 5, I will distinguish between two interpretations of the proposition that means need to be available to an agent. First is the internalist or cognitivist notion of availability, according to which available means are those about which the agent has instrumental beliefs. The second interpretation is external, according to which available means are those that are actually necessary for solving a task that the agent is performing.

In what follows, I will discuss what conditions need to be satisfied in order to be able to say that instrumental learning tasks show that psychopaths are instrumentally irrational. Furthermore, I will show how on both interpretations of availability, from the perspective of our current empirical knowledge it could be argued that Maibom (2005) does not conclusively show that psychopaths are more irrational than other people.

6.3 Internal rationality

In chapter 5, we saw that the cognitive system with level 2 rationality could be characterized in terms of beliefs, desires, intentions, and other propositional attitudes. The introduction of internal mental states provides space for applying conflicting criteria of rationality. To see why, let us distinguish between internal and external criteria of rationality. Rational criteria that apply to internal mental states are logically independent from criteria that apply to external performance. For example, if Mary orders gin and tonic, but unknowingly receives petroleum and drinks it, she is internally rational even though externally her action is not successful – that is, it does not satisfy her aim. On the other hand, a person who commits the gambler's fallacy but keeps winning money on the lottery could be seen as internally irrational, but externally rational – since her actions successfully accomplish her aims. Accordingly, principle (a) could be interpreted as applying to the internal functioning of mental states and their rational relations, or to external criteria that involve the success of the action in relation to the task that the cognitive agent is performing.

An internalist interpretation of (a) is provided by the principle R_1 (albeit in a slightly different terminology). According to R_1, "reason requires that if someone has an intrinsic desire

that p and a *belief that he can bring about p by bringing about q*, then he has an instrumental desire that he brings about q" (Smith, 2009, p. 119). Here the emphasis is on the belief; if an agent does not have the relevant instrumental belief about the means to her ends then she is not under the requirement R_1 to form an instrumental desire (see e.g. Kolodny & Brunero, 2013).

According to this internalist interpretation, for instance, Mary is not under the requirement of reason to form an instrumental desire to drink from a glass if she does not believe that drinking from the glass would satisfy some of her other desires. I will argue that similar reasoning applies to psychopaths since the dominant hypothesis that explain psychopaths' performance on instrumental learning tasks can plausibly be interpreted as indicating that they lack the relevant instrumental belief.[112]

When (a) is read internally, it can be maintained that, in instrumental learning tasks, psychopaths are not aware of the connection between the stimulus and punishment and do not detect changes in the reward/punishment contingency. There are two dominant hypotheses that pertain to explain why psychopaths perform worse than control groups on instrumental learning tasks: the integrated emotion systems hypothesis and the response modulation hypothesis (Moul, Killcross, & Dadds, 2012). I will start with the former.

James Blair's integrated emotion systems (IES) hypothesis explains the deficits of psychopaths in instrumental learning tasks in terms of impairments of the mechanisms that are involved in affectively targeting or representing certain stimuli (Blair, Mitchell, & Blair, 2005; Damasio, 1994). These mechanisms are responsible for categorizing or evaluating certain options or prospects and their outcomes as good or bad. By emotionally marking a certain type of stimulus as good or bad, the function of somatic markers is to establish associations between that type of stimuli and certain responses, thereby to induce a person to pursue a certain course of action or refrain from it. Thus, according to this explanation the problem of the gambling task, for example, is reduced to the problem of the inaccessibility of the representation that marks certain decks of cards as bad and others as good choices. On a more cognitive level, we could say that psychopaths lack the relevant representation that grounds instrumental belief about how to successfully solve the task.

Similarly the so-called response modulation hypothesis (RMH), advanced by Joseph Newman and his colleagues, predicts that the performance of psychopaths in instrumental learning tasks derives from the actual inaccessibility of some considerations, whether or not they are affectively marked (Hiatt & Newman, 2006; Koenigs & Newman, 2013). As supporters

[112] The argument provided in the next two paragraphs is based on (Jurjako & Malatesti, 2016).

of this view state, "behavior that characterizes criminal psychopaths results from failing to access information that nonpsychopathic individuals do access" (MacCoon & Newman, 2006, p. 803).

According to this account, psychopaths are less able to shift attention away from stimuli that are in their primary focus to secondary or contextual cues that are outside of their primary focus. On this account, the experiments with psychopaths on instrumental learning tasks, for example, show that they are impaired in directing attention to peripheral or contextual information that is pertinent to the task. For example, psychopaths do not automatically shift attention in response-reversal or gambling tasks because they do not *detect* the changes in the reinforcement contingencies (e.g. that the response to previously rewarded stimuli is now being punished). Again, we can say that the automatic mechanisms that are supposed to track cues from the environment do not produce relevant instrumental beliefs about how to solve the task.

Thus, according to the internalist interpretation of the availability clause in (a), we can say that the experiments do not show that psychopaths are more irrational than other people, because it is plausible that psychopaths lack the relevant instrumental belief about how to solve the task. In effect, we cannot say that psychopaths are irrational because they want to solve the task but fail to intend the means that they believe are necessary for solving the task. As the proposed explanations of psychopath's poor performance on these tasks show, they often lack the grounds for forming the relevant specific instrumental belief(s).

6.4 External rationality

Maibom's argument could be defended by assuming that the relevant notion of rationality is external. According to external rationality, the rationality of particular actions will depend on (1) the accuracy and, we might add, the presence of a relevant instrumental belief, and (2) the capacity of the instrumental belief in producing successful action (cf. Bermúdez, 2003, pp. 124-125). So far, we have been considering whether psychopaths are rational in the internal sense, in which (1) does not have to hold.

Thus, according to the external interpretation the argument could be that available means in requirement (a) should be construed as means that are available to normally functioning decision-making systems. As we have seen, psychopaths seem to have problems in affectively marking and/or in automatically directing attention to information that is often relevant for successfully solving a decision-making task (Moul, Killcross, & Dadds, 2012). Thus, judging by the criteria set by external rationality, since psychopaths either have inaccurate or

unavailable relevant instrumental beliefs, we could conclude that psychopaths are, in the external sense, more irrational than other people.

However, the plausibility of this argument might be doubted as well. To illustrate the first way in which it could be disputed, let us consider an example. If we were to ask a color-blind person to undertake a response-reversal task in which stimuli are colored it is likely that the person will perform worse than average. Even though color-blind people will perform worse than control groups on this type of task, we are not likely to say that they are more irrational than other people because of this. The reason for this opinion seems to be that the deficits lie in peripheral sensory systems that are not under the direct control of the agent, rather than in central cognitive systems that we usually associate with rationality. Similar things can be said about psychopaths' performance on instrumental learning tasks.

According to the response modulation and integrated emotion systems hypotheses, psychopaths exhibit deficits in peripheral sensory systems (automatic attention shifting and affective marking of behavioral alternatives, respectively) that process information at a subpersonal level, and not in higher-level cognitive systems that regulate full-blown belief production. In this respect, their deficits seem to be analogous to the case of a color-blind person. Since psychopaths do not have direct control of those systems it does not seem correct to say that they are less rational because of this.

It could be objected that the two cases are not analogous since color-blind people would perform normally if the instrumental learning task did not involve colored stimuli. However, one could respond that a similar thing could be said about psychopaths. Studies have shown that psychopaths' performance on instrumental learning tasks varies depending on the learning context (Brazil, et al., 2013; Moul, Killcross, & Dadds, 2012). For example, some studies have shown that psychopaths' performance on instrumental learning tasks is normalized when their attention is directed to relevant cues in the task (Hamilton, Hiatt Racer, & Newman, 2015; Koenigs & Newman, 2013). Other studies have failed to replicate the difference in behavioral performance in instrumental learning tasks even though they have confirmed that psychopaths' brains exhibit differential activation in regions that underlie instrumental learning tasks (Finger, et al., 2008; Gregory, et al., 2014). Thus, it seems that in this respect the analogy with color-blind people holds as well.

Regarding the second way in which the argument from external rationality can be disputed, we must remember that the criteria by which we can judge the rationality of decision-making processes and behavior can be both short- and long-term. Long-term criteria refer to adaptive behavior that directly influences the fitness of an agent. Short-term criteria refer to behaviors

and processes whose function is to promote reproductive and survival efforts in a more proximal way (see chapter 6 in Bermúdez, 2003). The requirement of external rationality, namely that one should have true instrumental beliefs in instrumental learning tasks, is also subject to the criteria of short- and long-term rationality. In addition, decision-making systems should be judged in relation to some 'design' features or psychological specifications of the psychological capacities of an agent and the environment in which agents with these decision-making systems process information and undertake actions (cf. Morton, 2010).

It might seem that psychopaths' frequent poor performance on instrumental learning tasks shows that their behavior is maladaptive and therefore externally irrational because it does not lead to successful behavior. This conclusion would be warranted only if it could be shown that psychopathic personality traits (1) lead to maladaptive behavior (2) because the 'design' features of the decision-making processes underlying those traits are defective given the particular tasks they have evolved to perform. However, if it could be shown that psychopathy represents just another *adaptive* life-strategy then it would not follow that psychopaths' frequent poor performance on a contextually limited set of learning tasks proves that they are externally more irrational then other people. Rather it could be argued that their performance in different learning tasks is an adaptation or a byproduct of an adaptation to a certain life-style.

In the rest of this chapter, I will provide evidence pertaining to show that psychopathy might be conceptualized as an adaptive life-history strategy. Furthermore, I will argue that psychopaths' contextually poor performances on instrumental learning tasks could plausibly be rationalized in the light of this evidence. The upshot of the present discussion should be that the burden of proof is on those who contend that psychopaths' performances on learning tasks show that they are instrumentally irrational, to argue either that psychopathy is not an adaptation or that despite being an adaptation psychopaths should still be considered instrumentally irrational.

6.4.1 Psychopathy, adaptation, and life-history strategies

At first glance, it might seem that being a psychopath is obviously maladaptive. Thomas Nadelhoffer and Walter Sinnott-Armstrong give a powerful description of psychopaths in order to show why this might be the case:

> Psychopaths are unable to lead normal, productive, healthy lives. One obvious reason is that they usually live much of their lives in prison, so they are deprived of freedom, which society values. Even when not incarcerated, they are often driven to lead nomadic lifestyles devoid

> of normal relationships with friends, family, and romantic partners. Love and friendship are also valued by society, and psychopaths are deprived of those benefits. Admittedly, psychopaths often see themselves not as deprived but rather as not restricted by useless attachments, but the lack of friendship and love still counts as 'harm or deprivation of benefit to the person as judged by the standards of the person's culture' (Wakefield 1992b: 384). Furthermore, owing to psychopaths' impaired ability to deliberate, form long-term goals, and adopt effective strategies for achieving these goals, their lives are often frustrating and deprived of achievements that are valued by society. Finally, and perhaps most important, psychopaths tend to die earlier than their nonpsychopathic counterparts, probably because of the correlations between psychopathy and violence, drug and alcohol abuse, and various kinds of risky behavior. (Nadelhoffer & Sinnot-Armstrong, 2013, pp. 244-245)

The authors rightly point out that it is easy to establish that psychopaths are deprived of *socially* valuable goods since those goods usually involve social and interpersonal relations. From this perspective their lives might seem to be "solitary, poor, nasty, brutish and short" (Hobbes, 1651/1985, p. 186). However, the acknowledgment that psychopaths are deprived of socially valuable goods does not do much to show that they are actually harming *themselves*, since their condition is characterized by pervasive, mostly intentional antisocial behavior (Blair, Mitchell, & Blair, 2005). Moreover, psychopaths usually do not see themselves as suffering from being what they are (Hare & Neumann, 2010; Nadelhoffer & Sinnot-Armstrong, 2013, pp. 246-247). The adaptationist thesis might explain why this could be the case.

The idea that psychopathy is an adaptation comes from a growing line of research that serves as a unified explanatory framework (Glenn & Raine, 2009) and a source of new research hypotheses (Krupp, Sewall, Lalumière, Sheriff, & Harris, 2012). In what follows, I will divide the case for seeing psychopathy as an evolutionary adaptation in three interrelated points: 1) core psychopathic traits are moderately to highly heritable; 2) antisocial behavior is adaptive in frequency-dependent sense; 3) psychopathic traits comprise traits that belong to one adaptive life history.

6.4.2 The heritability of psychopathic traits

Although the idea of genetic underpinnings of antisocial and criminal behavior has raised controversy among scholars, genetic research has provided increasing support for the assertion that significant part of the variance in antisocial behavior can be accounted for in terms of genetics (Moffitt, 2005; Raine, 2008). Similarly, when applied to psychopathy, genetic research shows that psychopathic traits are moderately to highly heritable (Glenn & Raine, 2014, p. 23). Initial twin studies have found that psychopathic traits are ~40%–50% heritable, where the

remainder of the variance is due to non-shared environmental influence (Blonigen, Carlson, Krueger, & Patrick, 2003; Blonigen, Hicks, Krueger, Patrick, & Iacono, 2005).[113] Heritability percentages related to psychopathic traits are consistent with percentages for other personality traits (Glenn & Raine, 2014, p. 23). However, the latter finding sets individuals with psychopathic traits apart from other general antisocial personality disorder, since twin studies have shown that, unlike other antisocial disorders, psychopathy is not correlated with a shared early environment.

Further studies found even higher heritability. In a large sample of twins (more than 7000 children) Viding and colleagues (Larsson, Viding, & Plomin, 2008; Viding, Blair, Moffitt, & Plomin, 2005; Viding, Jones, Frick, Moffitt, & Plomin, 2008) found that more than 60% of core psychopathic traits (those related to callousness and unemotionality) were heritable. In addition, they found that conduct disorder behaviors were more heritable (~75%) among children who possess core psychopathic traits than among those who did not.

It must be emphasized that a moderate to high heritability of psychopathic traits does not mean that psychopaths are genetically determined towards antisocial behavior.[114] As in all other similar cases proper explanation will involve genetic and environmental elements and their interaction (e.g. epigenetic factors). For example, it is widely thought that the psychopathic personality is stable over time and that this is mostly due to the genetic contribution. However, when changes do occur in psychopathic traits, these are mostly found to be influenced by environmental contributions (Glenn & Raine, 2014, p. 23). Moreover, environmental factors, such as growing up in abusive households, may play a role in aggravating antisocial features in psychopathic individuals (Blair R. J., 2007).

Genetic underpinnings of psychopathic traits may plausibly explain some noted neurobiological differences in psychopaths. James Blair (2007) has proposed that the

[113] Twin studies use data on dizygotic and monozygotic twins in order to determine how much phenotypic variability is due to genetics and how much is due to environmental influences. Roughly, the methodology is the following. Monozygotic twins share 100% of their genes, while dizygotic share 50% of their genes. Statistical calculations have shown that if the correlation between monozygotic twins on a given phenotypic trait is at least twice as big as the correlation between dizygotic twins, then we can conclude that the variance in the examined trait is due to genetic factors. When the difference in phenotypes is due to environmental influences, those influences can be due to shared and non-shared environments. Shared environments are usually the household in which the children grew up. Non-shared environments are environments that can be unique to children, such as schools and peer groups (see chapter 1 in Glenn & Raine, 2014).

[114] Here I just note that high heritability does not *necessarily* mean that the trait is genetically-determined (Sarkar, 1998; for a discussion of this issue see Sesardić, 2005).

hypoactivation of the amygdala in psychopathy is a consequence of genetic factors. This is important because it is widely held that abnormal functioning of the amygdala explains the abnormal performance of psychopaths in instrumental learning (especially response-reversal) tasks (Moul, Killcross, & Dadds, 2012). According to Blair, if psychopathy were a consequence of physical or mental abuse or neglect then we would expect, on the contrary, the amygdala to be hyperactive, as shown by animal studies and cases of PTSD in humans (cf. Blair R. J., 2007, p. 389).

The conjectured mechanisms that influence amygdala function include genes coding for serotonin transporter (SLC6A4). Nevertheless, in one study a correlation between one variant of that gene and increased callous–unemotional and narcissistic traits was found, but only when the relevant environmental factor included low socioeconomic status (cf. Glenn & Raine, 2014, p. 39). This shows how interaction between genetic and environmental factors can influence the display of psychopathic traits and provide a motive for some individuals to engage in antisocial behavior.

6.4.3 Adaptation and the frequency-dependence of antisocial behavior

Since core psychopathic traits are moderately to highly heritable we need an explanation for why those genes have remained in the human gene pool. One explanation could be that the section of the genome that is responsible for psychopathic traits is highly mutable, and that those mutations produce psychopathic traits at the phenotypic level (for a discussion see Glenn, Kurzban, & Raine, 2011). However, I contend that, for now, the biggest problem for the mutation hypothesis is that if psychopathy were a disorder like schizophrenia and autism, which are produced by gene mutations, we would expect those mutations to have negative fitness consequences, like they do in the case of schizophrenia and autism. In fact, there is no evidence that psychopaths' fitness is diminished (Krupp, Sewall, Lalumière, Sheriff, & Harris, 2013). The alternative adaptationist explanation plausibly shows why this might be the case. Moreover, the adaptationist explanation provides testable predictions that have been confirmed by empirical studies. In what follows I will try to provide theoretical framework and evidence in favor of thinking that psychopathy might be an adaptive, albeit morally highly dubious life strategy.

The most prominent adaptationist explanation of why genes coding for psychopathic traits were kept in the gene pool comes from game-theoretic considerations. Formal studies in game theory have shown that antisocial behavior can be adaptive when its prevalence in the general

population is low (Colman & Wilson, 1997; Harpending & Sobus, 1987; Mealey, 1995). The basic idea is that in a population that mostly consists of cooperative agents it becomes beneficial to play the cheater strategy, i.e. to take advantage of other agents' cooperative, helpful, or reciprocating dispositions without paying the cost of cooperation.[115] Being 'antisocial' in this formal sense is adaptive when the prevalence of non-cooperators or defectors is low enough not to be detected and retaliated against by the rest of the cooperative population. In this sense, antisocial strategies are maintained in a population by frequency-dependent selection processes. Colman and Wilson (1997) have estimated that the frequency of non-cooperators should be less than 2% in order for it to be an adaptive strategy, that is, to be included in an evolutionary equilibrium of behavioral strategies.[116]

When applied to psychopathy and translated into real-life evolutionary currency:

> these models presume that the general population is predominantly cooperative, honest, and trusting, which allows a small proportion of individuals to capitalize on this benevolence by cheating—stealing valuable resources and engaging in promiscuous sexual behavior (Mealey, 1995). As the proportion of cheaters (i.e., psychopaths) inches up in frequency, however, society at large becomes more vigilant, enacting counter-measures against their depredations (e.g., *imprisonment), thereby maintaining their frequency at a low level. (Skeem, Polaschek, Patrick, & Lilienfeld, 2012, p. 112)*

Incidentally, the estimated number of psychopaths in the general population is around 1% (Neumann & Hare, 2008). Although it is theoretical, this frequency-dependent model of selection explains why psychopathic traits, which predispose people to antisocial behavior, are constantly present in the population and relatively heritable.

Furthermore, studies in behavioral economy lend support to the idea that psychopathy is related to 'cheating' and other immoral behavioral strategies. In a study with incarcerated and non-incarcerated psychopaths, it was shown that they more often defect in one-shot and

[115] The term 'cheater strategy' refers to general antisocial behavior; it is not just a strategy that can be modeled by the prisoner's dilemma. Although some authors formally model antisocial behavior as a defection strategy in the prisoner's dilemma (Harpending & Sobus, 1987), others model it on the dove–hawk strategy (Colman & Wilson, 1997), where psychopaths are supposed to play the hawk strategy, which is more closely related to aggressive behavior. In empirical literature, there is evidence that the psychopathic personality is related to both the cheater and the hawk strategy (Book & Quinsey, 2004).

[116] It is worth noting that these formal models can be applied to genetic and social evolution (Colman & Wilson, 1997). So the idea that psychopathic traits are adaptations does not necessarily involve a commitment to having a strong genetic substrate.

repeated prisoner's dilemma games, respectively (Curry, Jones, & Viding, 2011; Mokros, et al., 2008).

Behavioral studies also reveal interesting underpinnings of brain activity among psychopaths. It has been found that in delivering moral judgments and making strategic decisions psychopaths employ brain areas that underpin emotionally *detached* processes (Joana B. Vieira, et al., 2013). While in normal subjects the decision to cooperate involves amygdala activity (automatic-emotional reaction) and the decision to defect involves greater activity in the dorsolateral prefrontal cortex (brain area associated with cognitive control that overrides prepotent impulses), in psychopathy the tendency of activation is reversed (Rilling, et al., 2007). This indicates how lower activation rates of the amygdala and other affect-related brain areas might enable psychopaths to play an antisocial strategy and in that sense to function as adapted by that particular behavioral strategy.

However, the idea that psychopathic traits and their neuropsychological underpinnings should not be considered as dysfunctions that enable irrational behavior but rather as traits that serve a particular life strategy is synthesized in a broader evolutionary-minded account of personality traits.

6.4.4 Psychopathy as a life-history strategy

As mentioned, most prevalent personality traits have a heritability of around 50% (Buss, 2009). The personality traits on which people differ include:

> individual differences in personality characteristics (e.g., dominance vs. submissiveness; agreeableness vs. aggressiveness), general intelligence and more specific abilities (e.g., spatial location vs. spatial rotation abilities), mating strategies (e.g., short term vs. long term), political attitudes (e.g., liberal vs. conservative), religiosity (high vs. low), body type (e.g., mesomorph, endomorph), mate value, and many others. *(Buss, 2009, p. 360)*

The perspective that evolutionary psychology brings to the fore emphasizes the importance that all these traits might have on "evolutionarily relevant outcomes, such as survival, mating success, offspring production, and parenting" (Buss, 2009, p. 360). In this respect, life-history theory has an especially illuminating role to play.

Life-history theory was developed in evolutionary ecology in order to explain how evolutionary adaptations to particular ecological niches produce diversity among the life histories of different species (Fabian & Flatt, 2012). One of the main concepts that emerged from this research is the trade-off between the time and energy that an organism must invest in

order to solve different environmental challenges related to its fitness (Stearns, 1989). Specifically, trade-offs must be made when an increase in fitness in one important life-history trait involves a fitness decrease in another important trait. For example, it has been found that there is a very basic negative genetic correlation between longevity and early reproduction. In particular, at the genetic level, laboratory studies have found that increased selection for extended lifespan causes a decrease in early reproduction among fruit flies, showing that there is an adaptive trade-off between longevity and the number of offspring (Fabian & Flatt, 2012).

When applied to personality traits, psychologists have delineated the following trade-offs from life-history theory (LH):

> (1) somatic effort (resources devoted to continued survival) versus reproductive effort (resources devoted to producing offspring),
>
> (2) parental effort versus mating effort,
>
> (3) quality versus quantity of offspring, and
>
> (4) future versus present reproduction. *(Glenn, Kurzban, & Raine, 2011, pp. 372-373)*

Applying these considerations to individuals involves thinking about trade-offs among resource allocations that these individuals have to make. For example:

> Energy can be allocated toward reproduction, which subsumes all of the effort required to successfully select, attract, and retain a mate, at least long enough for successful conception. Or energy can be allocated toward parenting and other forms of kin investment, which ultimately increase the reproductive success of genetic relatives. *(Buss, 2009, p. 361)*

What is relevant for the present discussion is that the four abovementioned dimensions of energy or resource trade-offs are correlated and form a continuum alongside which each individual could theoretically be placed (Glenn, Kurzban, & Raine, 2011). The extremes on the continuum form clusters of personality traits that are termed ‘slow’ and ‘fast’ LH strategy, and according to LH they tend to be selected together (Gladden, Figueredo, & Jacobs, 2009). Slow LH strategies involve extended longevity, delayed reproduction, greater investments in parental effort, and having a smaller number of offspring. Fast LH includes a shorter life span, early sexual activity, less investment in offspring, and greater reproduction.

The relevant supposition here is that psychopathic personality traits exemplify fast LH, which in turn supports traits relevant for maintaining frequency-dependent selection for the

'cheater' strategy (Glenn, Kurzban, & Raine, 2011; Krupp, Sewall, Lalumière, Sheriff, & Harris, 2013; Lalumière, Mishra, & Harris, 2008). In fact, psychopathy is considered to be a typical instance of fast LH strategy (Jonason, Koenig, & Tost, 2010; however see Gladden, Figueredo, & Jacobs, 2009).

An important source of evidence for this supposition comes from examining psychopathic traits as captured by PCL-R and other psychopathy measures and seeing how they fit to dimensions related to fast LH strategy. On the face of it, they seem to fit perfectly. For example, promiscuous sexual behavior and many short term marital relationships lead to reproductive and mating success; glibness/superficial charm enables psychopaths to navigate social structures undetected, present themselves as socially attractive, attract potential mates or poach the mates of others; impulsivity enables them to take advantage of immediate opportunities; being unemphatic and callous enables them to disregard potentially stressful stimuli, to lead careless life-styles, to disregard parental responsibilities, and carelessly pursue their often antisocial goals; etc. (Glenn, Kurzban, & Raine, 2011, p. 374). Studies have found correlations in all these aspects of psychopathy; the most salient from an evolutionary perspective are those that correlate psychopathy to short-term relationships and a higher number of sexual partners (Jonason, Li, & Buss, 2010).

The LH perspective enables us to recognize that "many apparent dysfunctions associated with psychopathy (e.g., reduced empathy, lack of guilt, impulsivity) may be better understood as design features of an extreme fast-spectrum strategy" (Del Giudice, 2014, p. 269). This perspective also sheds light on neuropsychological mechanisms in psychopathy. As mentioned, an under-activated amygdala seems to be necessary to reap the benefits of blunted affect and unempathic attitudes. Similarly, an under-activated ventromedial and orbitofrontal cortex may enable psychopaths to disregard the expected effects of punishment and to release the cognitive resources used for persisting in focused goal-directed behavior. Since this perseverance in immediate goal-directed action, related to a fast LH strategy, is the psychopath's default behavioral pattern, this might also explain why psychopaths must use top-down attentional resources and engage brain areas underlying cognitive control (e.g. the dorsal and ventrolateral areas of prefrontal cortex) in order to disengage from current goal-directed behavior and direct attention to other possibly relevant contextual stimuli (Hamilton, Hiatt Racer, & Newman, 2015).

The main point of the present discussion is that the observed abnormalities in psychopaths' neuropsychological capacities do not amount to functional failures in certain mechanisms when it comes to performing their naturally designed role, because the activity of these mechanisms

was shaped by natural selection. Reviewed considerations, including the heritability of psychopathic traits, the role those traits play in a fast life history strategy, and the frequency-dependent selection of antisocial traits strongly support the hypothesis that psychopathic symptomatology comprises "selected 'niche' adaptations" (Wakefield, 2000). This gives us a *prima facie* reason for thinking that psychopaths' performance on instrumental learning tasks does not indicate that they are instrumentally irrational. Rather, on the adaptationist interpretation, it just shows the (side-)effects of some decision-making processes that subserve a particular life history strategy.

6.4.5 Objection: Psychopathy and developmental mismatches

I argued that construing psychopathy as a particular life-history strategy undercuts the claim that psychopaths' decision-making processes are irrational because of abnormalities in the brain areas underpinning performance on instrumental learning tasks. However, one might argue that even if decision-making processes in psychopaths are adaptations to certain type of environments, in the present circumstances (Western, civilized, peaceful societies) they still present maladaptations. This would warrant us to describe their behavior as irrational. In other words, it could be argued that psychopathic traits, although they are adaptations to certain environments, are mismatched with fitness-relevant aspects of the current environment in which they live. To make the objection clear let us consider the following example.

Developmental plasticity allows organisms to adapt to changing ecological circumstances in which they find themselves. For example, a small crustacean *Daphnia cucullata* has the developmental ability to develop a helmet-shape head when raised in environments inhabited by predators. However, since the helmet-shaped head uses greater amounts of calories and limits the mobility of the *Daphnia*, in predator-free environments it does not develop such a defensive adaptation. The developmental mismatch happens when the trait that was adaptive in early environments becomes maladaptive in developmentally later environments (Garson, forthcoming; 2015). For example, if we grow *Daphnia* first in an environment where there are no predators and then place it in a different environment where there are predators, the developed traits will probably be maladaptive in the new environment.

Similarly, it could be argued that psychopathy is a developmental mismatch, and in this sense is a maladaptation that will probably lead to externally irrational behavior, even though psychopathic abnormalities could be completely functional with respect to their developmental 'design.' Relatedly, it is supposed that a fast LH strategy is adaptive in environments where

bioenergetical resources are low, and life and reproductive prospects are insecure (Gladden, Figueredo, & Jacobs, 2009). For example, growing up in a warzone or in an abusive household might trigger developmental adaptations that predict the future environment in which executing a fast LH might be reproductively advantageous. However, it might be the case that the future environment does not match what was 'predicted' in developmental conditions, thus making the traits that were adaptively functional (or rational) in the early environment maladaptive in the later environment. Translating this into the present context, it could be argued that psychopathy, even if it represents an adaptation to early environments (Pleistocene for humans), it is still a maladaptation that leads to irrational behavior because in the current environment a fast LH is not adaptive. Indeed, one could argue that psychopathy is externally irrational in this environment since according to some estimates, "93% of adult male psychopaths in the United States are in prison, jail, parole, or probation" (Kiehl & Hoffman, 2011, p. 355).

Nevertheless, the supposition that psychopathy represents a developmental mismatch is not well supported. First, even though psychopathy is related to lifelong entanglement with repressive state systems, this does not mean that psychopaths' fitness is reduced. As noted, a fast LH strategy predicts trade-offs between stronger investments in mating efforts as opposed to longevity. Furthermore, there is no evidence that psychopaths have reduced reproductive success (Krupp, Sewall, Lalumière, Sheriff, & Harris, 2013) and, as mentioned, in this respect they differ from people with mental disorders such as schizophrenia (Nadelhoffer & Sinnot-Armstrong, 2013).

Second, if psychopathy were a developmental mismatch this would imply that development of psychopathic traits involves conditional developmental rules of the following form: "if growing up in an abusive environment, develop antisocial traits," "if care-takers are emotionally detached, develop unempathic/callous traits," etc., which would reflect the early environment in which a person develops.

There are couple of reasons why psychopathy and its ensuing antisocial traits do not involve such developmental rules. If psychopathy were a conditional developmental strategy we could predict that "psychopaths have experienced difficult early conditions statistically predictive of an inhospitable future biotic or social environment or that they have reduced embodied capital and ability to compete" (Lalumière, Mishra, & Harris, 2008, p. 181). However, there is no evidence for this supposition. As mentioned, psychopathy is moderately to highly heritable and the environmental contribution comes from a non-shared environment (Glenn & Raine, 2014). In fact, there are some studies according to which, unlike other antisocial personality disorders, psychopathy is uncorrelated with ineffective parenthood (ibid.,

p. 182). Furthermore, psychopathy is not correlated with neurodevelopmental perturbations that would be indicative of the abovementioned unfavorable early environment (Harris, Rice, & Lalumière, 2001; Lalumière, Harris, & Rice, 2001). Thus, since it is implausible to suppose that psychopathy embodies a conditional developmental strategy it is *ipso facto* implausible to think that psychopaths will rigidly exhibit maladaptive and consequently irrational behavior in the developmental mismatch sense.

It might be objected that even though psychopathy might be an adaptation to past environments, executing a fast LH strategy is not adaptive in the present environment. However, if we grant that psychopathy is a result of frequency-dependent selection for 'cheater' strategies, then this objection loses its plausibility. As we have seen, there is no evidence that from the perspective of the fast LH strategy psychopaths' fitness is reduced. Furthermore, frequency-dependent selection predicts that antisocial behavior will be favorable if it is rare enough and if other cooperative agents do not have perfect memory and capacities for detecting and punishing 'cheaters' or antisocial agents (Colman & Wilson, 1997). These formal conditions plausibly apply to our current environment as well.

6.5 Concluding remarks

In this chapter, I have evaluated the argument according to which empirical data on instrumental learning tasks show that psychopaths are instrumentally irrational. One line of argument was to see whether psychopaths satisfy the requirement of instrumental rationality when the requirement is interpreted internally. I argued that, according to this interpretation, we are not justified in concluding that psychopaths do not satisfy the main requirement of instrumental rationality. Another line of argument was to see whether psychopaths satisfy the requirement of instrumental rationality when it is interpreted externally. Here I argued that things become a bit complicated because even though it can be argued that psychopaths have abnormal processes that underlie belief-formation mechanisms, we can still question whether the requirement of external rationality should be applied to psychopathy in general. In particular, I argued that if it could be shown that psychopathy is an adaptation to a certain life-style, then it could be claimed that psychopaths' decision-making processes are not externally irrational since their function is to enable this particular life-style. In this respect, I presented evidence that might indeed show that psychopathy is an adaptive, albeit highly immoral life-style strategy.

7 Conclusion

The main aim of this dissertation was, first, to investigate the nature of normative reasons, and then to assess the implications for our conceptions of normative reasons when we see them from the perspective of the methodological naturalism. In the second chapter, I followed Parfit (Parfit, 2011a) in distinguishing between object- and subject-based theories of normative reasons. There I argued that the naturalistic perspective seems to better align with subject-based theories of normative reasons. This commitment to subject-based theories set the direction and tone for the rest of the discussion in this thesis.

In the third chapter, I developed a version of a subject-based theory of normative reasons. I defended the idea that a response-dependence account of normative reasons can accommodate certain intuitive consideration that we connect with normative reasons. These considerations mainly involve the idea of *idealization* that we use when thinking about what reasons we have and what kind of reasons apply to us. I develop this account in analogy with response-dependentist theories of color ontology. In this respect, I argued that there is a good case to be made for thinking that the recognition of facts as reasons is partly determined by the cognitive/affective make up of people and their place in the world.

In the fourth chapter, I discussed normative reasons in the context of evolutionary debunking arguments. There I defended an argument according to which our best naturalistic understanding of normativity is incompatible with a strong mind-independent ontology of normative reasons. Thus, in that chapter, I gave a naturalistically based argument for the idea that a kind of subject-based theory of normative reasons should be endorsed, given the commitment to methodological naturalism.

In the fifth chapter, the goal was to further develop one type of subject-based theory of reasons. In particular, I explained the intuitive difference between categorical and hypothetical reasons and showed how their difference might be captured in terms of a subject-based theory of reasons. In order to do this, I developed a positive account of normative reasons, in which the notion of rationality plays a central role. I developed this account based on naturalistic explanations of how capacities of different cognitive complexity might have played, and might still play a role in determining what we think of as intuitively providing reasons for action. This view of reasons I interfaced with a model from game theory in order to show how categorical

reason-relations could have emerged, and how we can plausibly think about categorical reasons within a subject-based account of reasons.

In the sixth chapter, I apply the considerations developed in the previous chapters to a specific case study. The purpose of this chapter was to show how a naturalistically constrained account of normative reasons could be fruitfully applied in a practical context. In the contemporary literature there is an interesting debate about whether so-called unsuccessful or criminal psychopaths are rational or not. Among other things, the debate is interesting because the way in which we answer that question might have practical implications for the responsibility status of criminal psychopaths. For example, if it could be shown that psychopaths are practically irrational, then we would have a reason to deem them less than fully criminally responsible for their potential wrongdoings. So far, in the literature the most salient opinion seems to be that the empirical evidence shows that psychopaths are less then fully practically rational. In this chapter, I argued that this conclusion might be premature. I discussed the relevant evidence by interpreting it in accordance with internal and external construals of the norm of instrumental rationality. The conclusion of the chapter was that, given the two salient interpretations of the rationality condition, we have a reason to be suspicious of the relevant empirical evidence showing that psychopaths are less than fully rational. In this respect, we should be careful not to deliver too hasty conclusions on the responsibility status of psychopaths' given the currently available scientific data.

8 References

[1.] Aaltola, E. (2014). Affective Empathy as Core Moral Agency: Psychopathy, Autism and Reason Revisited. *Philosophical Explorations, 17*, 76-92.

[2.] Ainslie, G. (2001). *Breakdown of Will.* Cambridge: Cambridge University Press.

[3.] André, J.-B., & Morin, O. (2011). Questioning the Cultural Evolution of Altruism. *Journal of Evolutionary Biology, 24*, 2531-2542.

[4.] Anscombe, G. E. (1957). *Intention.* Oxford: Basil Blackwell.

[5.] Ardila, A. (2008). On the Evolutionary Origins of Executive Functions. *Brain and Cognition, 68*, 92–99.

[6.] Arkonovich, S. (2011). Advisors and Deliberation. *The Journal of Ethics, 15*, 405–424.

[7.] Axelrod, R. (1984). *The Evolution of Cooperation.* New York: Basic Books.

[8.] Babiak, P., & Hare, R. D. (2006). *Snakes in Suits: When Psychopaths Go to Work.* New York: Harper Collins Publishers Inc.

[9.] Baumard, N., André, J.-B., & Sperber, D. (2013). A Mutualistic Approach to Morality: The Evolution of Fairness by Partner Choice. *Behavioral and Brain Sciences, 36*, 59 – 122.

[10.] Beatty, J. (1995). The Evolutionary Contingency Thesis. In G. Wolters, & J. G. Lennox (Eds.), *Concepts, Theories, and Rationality in the Biological Sciences* (pp. 45-81). Konstanz: Universitätsverlag.

[11.] Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to Future Consequences Following Damage to Human Prefrontal Cortex. *Cognition, 50*, 7–15.

[12.] Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1999). Deciding Advantageously Before Knowing the Advantageous Strategy. *Science, 275*, 1293-1295.

[13.] Behrends, J. (2013). Meta-Normative Realism, Evolution and our Reasons to Survive. *Pacific Philosophical Quarterly, 94,* 486–502.

[14.] Bermúdez, J. L. (2003). *Thinking Without Words.* Oxford: Oxford University Press.

[15.] Bermúdez, J. L. (2005). *Philosophy of Psychology.* New York / Oxford: Routledge.

[16.] Binmore, K. (2007). *Game Theory: A Very Short Introduction.* Oxford: Oxford University Press.

[17.] Birch, J., & Okasha, S. (2015). Kin Selection and its Critics. *Bioscience, 65,* 22-32.

[18.] Blackburn, S. (1998). *Rulling Passions.* Oxford: Clarendon Press.

[19.] Blair, J. R., Mitchell, D. G., & Blair, K. (2005). *The Psychopath: Emotion and the Brain.* Oxford: Blackwell Publishing.

[20.] Blair, R. J. (2007). The Amygdala and Ventromedial Prefrontal Cortex in Morality and Psychopathy. *TRENDS in Cognitive Sciences, 11,* 387-392.

[21.] Blair, R. J. (2008). The Amygdala and Ventromedial Prefrontal Cortex: Functional Contributions and Dysfunction in Psychopathy. *Philosophical Transactions of the Royal Society, 363,* 2557–2565.

[22.] Blair, R. J., Colledge, E., & Mitchell, D. (2001). Somatic Markers and Response Reversal: is There Orbitofrontal Cortex Dysfunction in Boys with Psychopathic Tendencies? *Journal of Abnormal Psychology, 29,* 499–511.

[23.] Blonigen, D. M., Carlson, S. R., Krueger, R. F., & Patrick, C. J. (2003). A Twin Study of Self-reported Psychopathic Personality Traits. *Personality and Individual Differences, 35,* 179–197.

[24.] Blonigen, D. M., Hicks, B., Krueger, R., Patrick, C. J., & Iacono, W. (2005). Psychopathic Personality Traits: Heritability and Genetic Overlap with Internalizing and Externalizing Pathology. *Psychological Medicine, 25,* 637–648.

[25.] Book, A. S., & Quinsey, V. L. (2004). Psychopaths: Cheaters or Warrior-Hawks? *Personality and Individual Differences, 36,* 33–45.

[26.] Boudry, M., Blancke, S., & Pigliucci, M. (2015). What Makes Weird Beliefs Thrive? The Epidemiology of Pseudoscience. *Philosophical Psychology, 28*, 1177–1198.

[27.] Boyd, R., & Richerson, P. J. (1985). *Culture and the Evolutionary Process.* Chicago: University of Chicago Press.

[28.] Boyer, P. (2001). *Religion Explained: The Evolutionary Origins of Religious Thought.* New York: Basic Books.

[29.] Braddock, M. (2016). Evolutionary Debunking: Can Moral Realists Explain the Reliability of Our Moral Judgments? *Philosophical Psychology, Advanced Online Edition*, 1-14.

[30.] Brandt, R. (1979). *A Theory of the Good and the Right.* Oxford: Clarendon Press.

[31.] Bratman, M. (1996). Identification, Decision, and Treating as a Reason. *Philosophical Topics, 24*, 1-18.

[32.] Brazil, I. A., Maes, J. H., Scheper, I., Bulten, B. H., Kessels, R. P., Verkes, R. J., & A., d. B. (2013). Reversal Deficits in Individuals with Psychopathy in Explicit but not Implicit Learning Conditions. *Journal of Psychiatry and Neuroscience, 38*, E13-20.

[33.] Brembs, B. (2009). Mushroom Bodies Regulate Habit Formation in Drosophila. *Current Biology, 19*, 1351–1355.

[34.] Brink, D. O. (1997). Kantian Rationalism: Inescapability, Authority, and Supremacy. In G. Cullity, & B. Gaut (Eds.), *Ethics and Practical Reason* (pp. 255-292). Oxford: Oxford University Press.

[35.] Broome, J. (2004). Reasons. In J. Wallace, M. Smith, S. Scheffler, & P. Pettit (Eds.), *Reason and Value: Themes from the Moral Philosophy of Joseph Raz* (pp. 28–55). Oxford: Oxford University Press.

[36.] Broome, J. (2007). Is Rationality Normative? *Disputatio, 2*, 161-178.

[37.] Broome, J. (2013). *Rationality Through Reasoning.* Cambridge: Wiley Blackwell.

[38.] Brosnan, K. (2011). Do the Evolutionary Origins of our Moral Beliefs Undermine Moral Knowledge? *Biology and Philosophy*, 51-64.

[39.] Buss, D. M. (2009). How Can Evolutionary Psychology Successfully Explain Personality and Individual Differences? *Perspectives on Psychological Science, 4*, 359-366.

[40.] Byrne, A., & Hilbert, D. R. (2003). Color Realism and Color Science. *Behavioral and Brain Sciences, 26,* 3- 21.

[41.] Call, J., & Tomasello, M. (2008). Does the Chimpanzee Have a Theory of Mind? 30 Years Later. *Trends in Cognitive Science, 12,* 187–92.

[42.] Callebut, W. (2007). Herbert Simon's Silent Revolution. *Biological Theory, 2,* 76–86.

[43.] Clark, A. (2000). *A Theory of Sentience.* Oxford: Oxford University Press.

[44.] Colman, A. M., & Wilson, C. J. (1997). Antisocial Personality Disorder: An Evolutionary Game Theory Analysis. *Legal and Criminological Psychology, 2,* 23-34.

[45.] Colyvan, M. (2009). Naturalizing Normativity. In D. Braddon-Mitchell, & R. Nola (Eds.), *Conceptual Analysis and Philosophical Naturalism* (pp. 303-313). Cambridge: A Bradford book, MIT Press.

[46.] Cooke, D. J. (1998). Psychopathy Across Cultures. In D. J. Cooke, A. E. Forth, & R. D. Hare (Eds.), *Psychopathy: Theory, Research and Implications for Society* (pp. 13-46). Dordrecht: Springer Science+Business Media.

[47.] Copp, D. (1995). *Morality, Normativity, and Society.* Oxford: Oxford University Press.

[48.] Cuneo, T. (2007). *The Normative Web.* Oxford: Oxford University Press.

[49.] Curry, O., Jones, M. C., & Viding, E. (2011). The Psychopath's Dilemma: The Effects of Psychopathic Personality Traits in One-shot Games. *Personality and Individual Differences, 50,* 804–809.

[50.] Curtis, V., Aunger, R., & Rabie, T. (2004). Evidence that Disgust Evolved to Protect from Risk of Disease. *Proceedings of the Royal Society of London, Series B, 271*, S131–S133.

[51.] Damasio, A. (1994). *Descartes' Error.* New York: Putnam Publishing.

[52.] Dancy, J. (2004). *Ethics Without Principles.* Oxford: Oxford University Press.

[53.] Daniels, N. (1996). *Justice and Justification: Reflective Equilibrium in Theory and Practice.* New York: Cambridge University Press.

[54.] Davidson, D. (2001). *Essays on Actions and Events.* Oxford: Oxford University Press.

[55.] Dawkins, R. (1986). *Unravelling Animal Behaviour.* Harlow, UK: Longman.

[56.] de Lazari-Radek, K., & Singer, P. (2012). The Objectivity of Ethics and the Unity of Practical Reason. *Ethics, 123*(1), 9-31.

[57.] de Waal, F. (1996). *Good Natured: The Origins of Right and Wrong in Humans and Other Animals.* Cambridge: Harvard University Press.

[58.] Deem, M. J. (2016). Dehorning the Darwinian Dilemma for Normative Realism. *Biology and Philosophy, Advanced Online Edition*, 1-20.

[59.] Del Giudice, M. (2014). An Evolutionary Life History Framework for Psychopathology. *Psychological Inquiry, 25*, 261–300.

[60.] Dennett, D. C. (2003). *Freedom Evolves.* London: Penguin Books.

[61.] Doris, J. M., & Stich, S. P. (2012). As a Matter of Fact: Empirical Perspectives on Ethics. In S. Stich, *Collected Papers: Knowledge, Rationality, and Morality* (Vol. II, pp. 114-152). Oxford: Oxford University Press.

[62.] El Mouden, C., André, J.-B., Morin, O., & Netlle, D. (2013). Cultural Transmission and the Evolution of Human Behaviour: a General Approach Based on the Price Equation. *Journal of Evolutionary Biology*, 231-241.

[63.] El Mouden, C., Burton-Chellew, M., Gardner, A., & West, S. (2012). What do Humans Maximise? In S. Okasha, & K. Binmore (Eds.), *Evolution and*

Rationality: Decisions, Cooperation and Strategic Behaviour (pp. 23-49). Cambridge: Cambridge University Press.

[64.] Enoch, D. (2005). Why Idealize? *Ethics, 115*, 759-787.

[65.] Enoch, D. (2010). The Epistemological Challenge to Metanormative Realism: How Best to Understand it, and How to Cope with it. *Philosophical studies, 148*, 413-438.

[66.] Enoch, D. (2011). *Taking Morality Seriously.* Oxford: Oxford University Press.

[67.] Fabian, D., & Flatt, T. (2012). Life History Evolution. *Nature Education Knowledge, 3*.

[68.] Fehr, E., & Fischbacher, U. (2003). The Nature of Human Altruism. *Nature, 425*, 785-791.

[69.] Fehr, E., & Gächter, S. (2000). Cooperation and Punishment in Public Goods Experiments. *American Economic Review, 90*, 980-994.

[70.] Fehr, E., & Gächter, S. (2002). Altruistic Punishment in Humans. *Nature, 412*, 137-140.

[71.] Finger, E. C., Marsh, A. A., Mitchell, D. G., Reid, M. E., Sims, C., Budhani, S., . . . Blair, R. J. (2008). Abnormal Ventromedial Prefrontal Cortex Function in Children with Psychopathic Traits during Reversal Learning. *Archives of General Psychiatry, 65*, 586–594.

[72.] Finlay, S. (2006). The Reasons that Matter. *Australasian Journal of Philosophy, 84*, 1 – 20.

[73.] Finlay, S. (2009). Oughts and Ends. *Philosophical studies, 143*, 315–340.

[74.] Finlay, S. (2010a). What Ought Probably Means, and Why you Can't Detach it. *Synthese, 177*, 67–89.

[75.] Finlay, S. (2010b). Recent Work on Normativity. *Analysis, 70*, 331-346.

[76.] Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

[77.] FitzPatrick, W. J. (2008). Robust Ethical Realism, Non-Naturalism, and Normativity. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol. 3, pp. 159-205). Oxford: Oxford Unversity Press.

[78.] Focquaert, F., Glenn, A., & Raine, A. (2015). Psychopathy and Free Will from a Philosophical and Cognitive Neuroscience Perspective. In W. Glannon (Ed.), *Free Will and the Brain* (pp. 103-124). Cambridge:: Cambridge University Press.

[79.] Frankfurt, H. (1988a). The Importance of What We Care About. In H. Frankfurt, *The Importance of What We Care About* (pp. 80-94). Cambridge: Cambridge University Press.

[80.] Frankfurt, H. (1988b). Identification and Wholeheartedness. In H. Frankfurt, *The Importance of What We Care About* (pp. 159-176). Cambridge: Cambridge University Press.

[81.] Gao, Y., & Raine, A. (2010). Successful and Unsuccessful Psychopaths: A Neurobiological Model. *Behavioral Sciences & the Law, 28,* 194–210.

[82.] Garson, J. (2015). *The Biological Mind: A Philosophical Introduction.* London and New York: Routledge: Taylor and Francis Group.

[83.] Garson, J. (forthcoming). The Developmental Plasticity Challenge to Wakefield's View. In L. Faucher, & D. Forest (Eds.), *Defining Mental Disorder: Jerome Wakefield and his Critics.* Cambridge, MA: MIT Press.

[84.] Gaus, G. (2011). *The Order of Public Reason.* Cambridge: Cambridge University Press.

[85.] Gauthier, D. (1986). *Morals by Agreement.* Oxford: Oxford University Press.

[86.] Gettier, E. (1963). Is Justified True Belief Knowledge? *Analysis, 23,* 121-123.

[87.] Gibbard, A. (1990). *Wise Choices, Apt Feelings.* Cambridge: Harvard University Press.

[88.] Gibson, J. J. (1979). *The Ecological Approach to Visual Perception.* Boston: Houghton Mifflin.

[89.] Giere, R. N. (2006). *Scientific Perspectivism.* Chicago and London: University of Chicago Press.

[90.] Giere, R. N. (2008). Naturalism. In S. Psillos, & M. Curd (Eds.), *The Routledge Companion to Philosophy of Science* (pp. 213-223). London and New York: Routledge.

[91.] Gintis, H. (2003). The hitchhiker's Guide to Altruism: Gene-culture Coevolution, and the Internalization of Norms. *Journal of Theoretical Biology, 220,* 407–418.

[92.] Gladden, P. R., Figueredo, A. J., & Jacobs, J. W. (2009). Life History Strategy, Psychopathic Attitudes, Personality, and General Intelligence. *Personality and Individual Differences, 46,* 270–275.

[93.] Glenn, A. L., & Raine, A. (2009). Psychopathy and Instrumental Aggression: Evolutionary, Neurobiological, and Legal Perspectives. *International Journal of Law and Psychiatry, 32,* 253–258.

[94.] Glenn, A. L., & Raine, A. (2014). *Psychopathy: An Introduction to Biological Findings and Their Implications.* New York and London: New York University Press.

[95.] Glenn, A. L., Kurzban, R., & Raine, A. (2011). Evolutionary Theory and Psychopathy. *Aggression and Violent Behavior, 16,* 371-380.

[96.] Goldman, A. (2010). *Reasons from Within: Desires and Values.* Oxford: Oxford University Press.

[97.] Greene, J., Sommerville, B. R., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science, 293,* 2105-2108.

[98.] Gregory, S., Blair, R. J., Ffytche, D., Simmons, A., Kumari, V., Hodgins, S., & Blackwood, N. (2014). Punishment and Psychopathy: a Case-control Functional MRI Investigation of Reinforcement Learning in Violent Antisocial Personality Disordered Men. *Lancet Psychiatry, 2,* 153–160.

[99.] Griesemer, J., & Szathmáry, E. (2009). Gánti's Chemoton Model and Life Criteria. In S. Rasmussen, M. A. Bedau, C. Liaohai, D. Deamer, D. C. Krakauer, N. H. Packard, & P. F. Stadler (Eds.), *Protocells: Bridging Nonliving and Living Matter* (pp. 481-512). Cambridge, Mass.: The MIT Press.

[100.] Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Model of Moral Judgment. *Psychological review, 108*, 814-834.

[101.] Haidt, J. (2007). The New Synthesis in Moral Psychology. *Science, 316*, 998-1002.

[102.] Haidt, J., & Bjorklund, F. (2008). Social Intuitionists Answer Six Questions about Moral Psychology. In W. Sinnott-Armstrong (Ed.), *Moral Psychology: The Cognitive Science of Morality: Intuition and Diversity* (Vol. II, pp. 181-217). Cambirdge, Mass.: A Bradford Book, The MIT Press.

[103.] Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, Culture, and Morality, or is it Wrong to Eat your Dog? *Journal of Personality and Social Psychology, 65*, 613-628.

[104.] Hall, J. R., & Benning, S. D. (2006). The 'Successful' Psychopath: Adaptive and Subclinical Manifestations of Psychopathy in the General Population. In C. J. Patrick (Ed.), *Handbook of Psychopathy* (pp. 459–478). New York, London: The Guildford Press.

[105.] Hamilton, R. K., Hiatt Racer, K., & Newman, J. P. (2015). Impaired Integration in Psychopathy: A Unified Theory of Psychopathic Dysfunction. *Psychological Review, 122*, 770–791.

[106.] Hamilton, W. D. (1964). The Genetical Evolution of Social Behaviour, I, II. *Journal of Theoretical Biology, 7*, 1-52.

[107.] Hardin, C. L. (1988). *Color for Philosophers: Unweaving the Rainbow.* Indianapolis/Cambridge: Hackett Publishing.

[108.] Hardin, C. L. (2003). A Reflectance Doth Not a Color Make. *The Journal of Philosophy*, 191-202.

[109.] Hare, R. D. (1993). *Without Conscience: The Disturbing World of Psychopaths Among Us.* New York: Guilford Press.

[110.] Hare, R. D. (2003). *The Psychopathy Checklist-Revised.* Toronto: Multi-Health Systems.

[111.] Hare, R. D., & Neumann, C. S. (2010). Psychopathy: Assessment and Forensic Implications. In L. Malatesti, & J. McMillan (Eds.), *Responsibility and Psychopathy: Interfacing Law, Psychiatry and Philosophy* (pp. 93-124). Oxford: Oxford University Press.

[112.] Harman, G. (2000). Is There a Single True Morality. In G. Harman, *Explaining Value and Other Essays in Moral Philosophy* (pp. 77-99). Oxford: Clarendon Press.

[113.] Harman, G. (2004). Practical Aspects of Theoretical Reasoning. In A. R. Mele, & P. Rawling (Eds.), *The Oxford Handbook of Rationality.* (pp. 45-56). Oxford: Oxford University Press.

[114.] Harms, W. (2004). *Information and Meaning in Evolutionary Processes.* Cambridge: Cambridge University Press.

[115.] Harms, W., & Skyrms, B. (2008). Evolution of Moral Norms. In M. Ruse (Ed.), *The Oxford Handbook of Philosophy of Biology* (pp. 434-450). Oxford: Oxford University Press.

[116.] Harpending, H. C., & Sobus, J. (1987). Sociopathy as an Adaptation. *Ethology and Sociobiology, 8,* 63S-72s.

[117.] Harris, G. T., Rice, M. E., & Lalumière, M. (2001). Criminal Violence: The Roles of Psychopathy, Neurodevelopmental Insults, and Antisocial Parenting. *Criminal Justice and Behavior, 28*, 402-426.

[118.] Hauser, M. D., Young, L., & Cushman, F. (2008). Reviving Rawls's Linguistic Analogy: Operative Principles and the Causal Structure of Moral Actions. In W. Sinnott-Armstrong (Ed.), *Moral Psychology: The Cognitive Science of Morality: Intuition and Diversity* (Vol. II, pp. 107-143). Cambridge, Mass: A Bradford book, The MIT Press.

[119.] Heatwood, C. (2005). The Problem of Defective Desires. *Australasian Journal of Philosophy, 83*, 487-504.

[120.] Hiatt, K. D., & Newman, J. P. (2006). Understanding Psychopathy: The Cognitive Side. In C. Patrick (Ed.), *Handbook of Psychopathy* (pp. 334–352). New York: Guildford Press.

[121.] Hobbes, T. (1651/1985). *Leviathan* (10th ed.). (C. B. MacPherson, Ed.) London: Penguin Books.

[122.] Hooker, B., & Streumer, B. (2004). Procedural and Substantive Practical Rationality. In A. R. Mele, & P. Rawling (Eds.), *Oxford Handbook of Rationality* (pp. 57-74). Oxford: Oxford University Press.

[123.] Huemer, M. (2005). *Ethical Intuitionism.* New York: Palgrave Macmillan.

[124.] Huttegger, S. (2007). Evolutionary Explanations of Indicatives and Imperatives. *Erkenntnis, 66*, 409–436.

[125.] Jackson, F. (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis.* Oxford: Oxford University Press.

[126.] Joana B. Vieira, J. B., Almeida, P. R., Ferreira-Santos, F., Barbosa, F., Marques-Teixeira, J., & Marsh, A. A. (2013). Distinct Neural Activation Patterns Underlie Economic Decisions in High and Low Psychopathy Scorers. *Social Cognitive and Affective Neuroscience*, 1-9.

[127.] Johnston, M. (1989). Dispositional Theories of Value. *Proceedings of the Aristotelian Society, Supplementary Volumes, 63*, 139-174.

[128.] Jonason, P. K., Koenig, B. L., & Tost, J. (2010). Living a Fast Life: The Dark Triad and Life History Theory. *Human Nature, 21*, 428-442.

[129.] Jonason, P. K., Li, N. P., & Buss, D. M. (2010). The Costs and Benefits of the Dark Triad: Implications for Mate Poaching and Mate Retention Tactics. *Personality and Individual Differences, 48*, 373-378.

[130.] Joyce, R. (2001). *The Myth of Morality.* Cambridge: Cambridge University Press.

[131.] Joyce, R. (2006). *The Evolution of Morality.* Cambridge, MA: MIT Press.

[132.] Joyce, R. (2013). The Evolutionary Debunking of Morality. In J. Feinberg, & R. Shafer-Landau (Eds.), *Reason and Responsibility: Readings in Some Basic Problems of Philosophy* (pp. 527-534). Boston: Wadsworth Cenage Learning.

[133.] Jurjako, M., & Malatesti, L. (2016). Instrumental Rationality in Psychopathy: Implications from Learning Tasks. *Philosophical Psychology,* Online edition.

[134.] Kacelnik, A. (2006). Meanings of Rationality. In M. Nudds, & S. Hurley (Eds.), *Rational Animals?* (pp. 87-106). Oxford: Oxford University Press.

[135.] Kahane, G. (2011). Evolutionary Debunking Arguments. *Noûs, 45*, 103-125.

[136.] Kennett, J. (2010). Reasons, Emotion, and Moral Judgment in the Psychopath. In L. Malatesti, & J. McMillan (Eds.), *Responsibility and Psychopathy: Interfacing Law, Psychiatry and Philosophy* (pp. 243-260). Oxford: Oxford University Press.

[137.] Kennett, J., & Fine, C. (2009). Will the Real Moral Judgment Please Stand Up? *Ethical Theory and Moral Practice, 12*, 77–96.

[138.] Kiehl, K. A. (2008). Without Morals: The Cognitive Neuroscience of Criminal Psychopaths. In W. Sinnot-Armstrong (Ed.), *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development,* (Vol. III, pp. 166-171). Cambridge, Mass.: MIT Press.

[139.] Kiehl, K. A., & Hoffman, M. B. (2011). The Criminal Psychopath: History, Neuroscience, and Economics. *Jurimetrics, 51*, 355-397.

[140.] Koenigs, M., & Newman, J. P. (2013). The Decision Making Impairment in Psychopathy: Psychological and Neurobiological Mechanisms. In W. P. Sinnott-Armstrong, & K. A. Kiehl (Eds.), *Handbook on Psychopathy and Law* (pp. 93-106). Oxford: Oxford University Press.

[141.] Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgments. *Nature, 446*, 908–911.

[142.] Kolodny, N. (2005). Why Be Rational? *Mind, 114*, 509-563.

[143.] Kolodny, N., & Brunero, J. (2013, February 13). *Instrumental Rationality.* Retrieved April 5, 2015, from The Stanford Encyclopedia of Philosophy: http://plato.stanford.edu/archives/fall2013/entries/rationality-instrumental/

[144.] Korsgaard, C. (1986). Scepticism about Practical Reason. *The Journal of Philosophy, 83*, 5-25.

[145.] Korsgaard, C. (1996). *The Sources of Normativity.* Cambridge: Cambridge University Press.

[146.] Korsgaard, C. (2009). *The Constitution of Agency: Essays on Practical Reason and Moral Psychology.* Oxford: Oxford University Press.

[147.] Korsgaard, C. (2011). The Activity of Reason. In R. J. Wallace, R. Kumar, & S. Freeman (Eds.), *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon* (pp. 3-22). Oxford: Oxford University Press.

[148.] Krebs, D. L. (2011). *The Origins of Morality: An Evolutionary Account.* New York: Oxford University Press.

[149.] Krupp, D. B., Sewall, L. A., Lalumière, M. L., Sheriff, C., & Harris, G. T. (2012). Nepotistic Patterns of Violent Psychopathy: Evidence for Adaptation? *Frontiers in Psychology, 3*, 1-8.

[150.] Krupp, D. B., Sewall, L. A., Lalumière, M. L., Sheriff, C., & Harris, G. T. (2013). Psychopathy, Adaptation, and Disorder. *Frontiers in Psychology, 4*, 1-5.

[151.] Laland, K. N. (2008). Exploring Gene–culture Interactions: Insights from Handedness, Sexual Selection and Niche-construction Case Studies. *Philosophical Transactions of the Royal Society B, 363*, 3577–3589.

[152.] Lalumière, M., Harris, G. T., & Rice, M. E. (2001). Psychopathy and Developmental Instability. *Evolution and Human Behavior, 22*, 75-92.

[153.] Lalumière, M., Mishra, S., & Harris, G. T. (2008). In Cold Blood: The Evolution of Psychopathy. In J. Duntley, & T. J. Shackelford (Eds.), *Evolutionary Forensic Psychology* (pp. 176-197). Oxford: Oxford University Press.

[154.] Larsson, H., Viding, E., & Plomin, R. (2008). Callous-unemotional Traits and Antisocial Behavior: Genetic, Environmental, and Early Parenting Characteristics. *Criminal Justice and Behavior, 35*, 197–211.

[155.] Lenman, J. (2009). Reasons for Action: Justification vs. Explanation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2011 ed.). Retrieved from http://plato.stanford.edu/archives/win2011/entries/reasons-just-vs-expl/

[156.] Lewis, D. (1969). *Convention.* Cambridge, Mass.: Harvard University Press.

[157.] Lewis, D. (1989). Dispositional Theories of Value. *Proceedings of the Aristotelian Society*, 113–137.

[158.] Locke, J. (1690/1995). *An Essay Concerning Human Understanding.* New York: Prometheus Books.

[159.] MacCoon, D. G., & Newman, J. P. (2006). Content Meets Process: Using Attributions and Standards to Inform Cognitive Vulnerability in Psychopathy, Antisocial Personality Disorder, and Depression. *Journal of Social and Clinical Psychology, 25*, 802-824.

[160.] Mackie, J. (1977/1990). *Ethics: Inventing Right and Wrong.* London: Penguin books.

[161.] Maibom, H. (2005). Moral Unreason: The Case of Psychopathy. *Mind and Language, 20*, 237–257.

[162.] Maibom, H. (2010). Rationalism, Emotivism, and the Psychopath. In L. Malatesti, & J. McMillan (Eds.), *Responsibility and Psychopathy: Interfacing Law, Psychiatry and Philosophy* (pp. 227-241). Oxford: Oxford University Press.

[163.] McDowell, J. (1985). Values and Secondary Qualities. In T. Honderich (Ed.), *Objectivity and Morality* (pp. 110-129). London: Routledge.

[164.] Mealey, L. (1995). The Sociobiology of Sociopathy: An Integrated Evolutionary Model. *Behavioral and Brain Sciences, 18*, 460-485.

[165.] Mercier, H., & Sperber, D. (2011). Why Do Humans Reason? Arguments for an Argumentative Theory. *Behavioral and Brain Sciences, 34*, 57–111.

[166.] Millgram, E. (1995). Was Hume a Humean? *Hume Studies, 21*, 75-93.

[167.] Millikan, R. G. (1989). Biosemantics. *Journal of Philosophy, 86*, 288–302.

[168.] Millikan, R. G. (1996). Pushmi-Pullyu Representations. In L. May, M. Friedman, & A. Clark (Eds.), *Mind and Morals: Essays on Cognitive Science and Ethics* (pp. 145–161). Cambridge, Mass.: MIT Press.

[169.] Milo, R. (1995). Contractarian Constructivism. *Journal of Philosophy, 92*, 181-204.

[170.] Miščević, N. (2006). Moral Concepts: From Thickness to Response-Dependence. *Acta Analytica, 21*, 3-32.

[171.] Miščević, N. (2011). No More Tears in Heaven: Two Views of Response-Dependence. *Acta Analytica, 26*, 75-93.

[172.] Miščević, N. (2012). Color. *Croatian Journal of Philosophy, 12*(3), 489-507.

[173.] Mitchell, D. G., Colledge, E., Leonard, A., & Blair, R. J. (2002). Risky Decisions and Response Reversal: Is There Evidence of Orbitofrontal Cortex Dysfunction in Psychopathic Individuals? *Neuropsychologia,, 40*, 2013–2022.

[174.] Moffitt, T. E. (2005). The New Look of Behavioral Genetics in Developmental Psychopathology: Gene-environment Interplay in Antisocial Behaviors. *Psychological Bulletin, 131*, 533–554.

[175.] Mokros, A., Menner, B., Eisenbarth, H., Alpers, G. W., Lange, K. W., & Osterheider, M. (2008). Diminished Cooperativeness of Psychopaths in a Prisoner's Dilemma Game Yields Higher Rewards. *Journal of Abnormal Psychology, 117*, 406–413.

[176.] Mollon, J. D. (1989). "Tho' she kneel'd in that place where they grew.." The Uses and Origins of Primate Colour Vision. *The Journal of Experimental Biology, 146*, 21–38.

[177.] Morin, O. (2014). Is Cooperation a Maladaptive By-product of Social Learning? The Docility Hypothesis Reconsidered. *Theoretical Biology, 9*, 286–295.

[178.] Morse, S. J. (2000). Rationality and Responsibility. *Southern California Law Review, 74*, 251-268.

[179.] Morton, J. M. (2010). Toward an Ecological Theory of the Norms of Practical Deliberation. *European Journal of Philosophy*, 561-584.

[180.] Moul, C., Killcross, S., & Dadds, M. R. (2012). A Model of Differential Amygdala Activation in Psychopathy. *Psychological Review, 119*, 789-806.

[181.] Nadelhoffer, T., & Sinnot-Armstrong, W. (2013). Is Psychopathy a Mental Disease? In N. Vincent (Ed.), *Neuroscience and Legal Responsibility* (pp. 229-255). Oxford: Oxford University Press.

[182.] Neumann, C. S., & Hare, R. D. (2008). Psychopathic Traits in a Large Community Sample: Links to Violence, Alcohol Use, and Intelligence. *Journal of Consulting and Clinical Psychology, 76*, 983-899.

[183.] Nichols, S. (2004). *Sentimental Rules: On the Natural Foundation of Moral Judgement.* Oxford: Oxford University Press.

[184.] Nisbett, R., & Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgment.* Englewood Cliffs: Prentice-Hall.

[185.] Nowak, M. A. (2006). Five Rules for the Evolution of Cooperation. *Science, 314*, 1560-1563.

[186.] Nowak, M. A., Tarnita, C. E., & Wilson, E. O. (2010). The Evolution of Eusociality. *Nature, 466*(7310), 1057–1062.

[187.] Okasha, S. (2006). *Evolution and the Levels of Selection.* Oxford: Clarendon Press.

[188.] Olson, J. (2014). *Moral Error Theory: History, Critique, Defence.* Oxford: Oxford University Press.

[189.] O'Neill, O. (1996). Introduction. In C. Korsgaard, *The Sources of Normativity* (pp. xi-xv). Oxford: Oxford University Press.

[190.] O'Neill, O. (2001). Consistency in Action. In E. Millgram (Ed.), *Varieties of Practical Reasoning* (pp. 301-329). Cambridge MA.: MIT Press.

[191.] Osumi, T., & Ohira, H. (2010). The Positive Side of Psychopathy: Emotional Detachment in Psychopathy and Rational Decision-making in the Ultimatum Game. *Personality and Individual Differences, 49*, 451-456.

[192.] Palmer, S. (2000). *Vision Science: Photons to Phenomenology.* Cambridge, Mass.: MIT Press.

[193.] Papineau, D. (2004). *Thinking about Consciousness.* Oxford: Oxford University Press.

[194.] Papineau, D. (2009). Naturalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Retrieved from The Stanford Encyclopedia of Philosophy.

[195.] Parfit, D. (1984). *Reasons and Persons.* Oxford: Oxford University Press.

[196.] Parfit, D. (1997). Reasons and Motivation. *Proceedings of the Aristotelian Society, Supplementary Volumes, 71*, 99-146.

[197.] Parfit, D. (2006). Normativity. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol. I, pp. 325-380). Oxford: Clarendon Press.

[198.] Parfit, D. (2011a). *On What Matters* (Vol. I). Oxford: Oxford University Press.

[199.] Parfit, D. (2011b). *On What Matters* (Vol. II). Oxford: Oxford University Press.

[200.] Parfit, D. (forthcoming). *On What Matters* (Vol. III). Oxford: Oxford University Press.

[201.] Peacocke, C. (1992). *A Study of Concepts.* Cambridge, Mass.: MIT Press.

[202.] Pollock, J. L. (1987). Defeasible Reasoning. *Cognitive Science*, 481-518.

[203.] Prinz, J. (2006). The Emotional Basis of Moral Judgment. *Philosophical Explorations*, 29–43.

[204.] Quine, W. v. (1981). *Theories and Things.* Cambridge: Harvard University Press.

[205.] Quinn, W. (1993). Putting Rationality in its Place. In W. Quinn, *Morality and Action.* Cambridge: Cambridge University Press.

[206.] Railton, P. (1986). Moral Realism. *The Philosophical Review, 95*, 163-207.

[207.] Railton, P. (2004). How to Engage Reason: The Problem of Regress. In R. J. Wallace, P. Pettit, S. Scheffler, & M. Smith (Eds.), *Reason and Value* (pp. 176-201). New York: Oxford University Press.

[208.] Railton, P. (2009). Practical Competence and Fluent Agency. In D. Sobel, & S. Wall (Eds.), *Reasons for Action* (pp. 81-115). Cambridge: Cambridge University Press.

[209.] Raine, A. (2008). From Genes to Brain to Antisocial Behavior. *Current Directions in Psychological Science, 17*, 323–328.

[210.] Ramirez, E. (2015). Receptivity, Reactivity and the Successful Psychopath. *Philosophical Explorations, 18*, 330-343.

[211.] Ravenscroft, I. (2010). Folk Psychology as a Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

[212.] Rawls, J. (1971). *A Theory of Justice.* Harvard: Harvard University Press.

[213.] Raz, J. (1975). *Practical Reason and Norms.* Oxford: Oxford University Press.

[214.] Raz, J. (1999). *Engaging Reason.* Oxford: Oxford University Press.

[215.] Reisner, A., & Steglich-Petersen, A. (Eds.). (2011). *Reasons for Belief.* Cambridge: Cambridge University Press.

[216.] Richerson, P. J., & Boyd, R. (2005). *Not by Genes Alone: How Culture Transformed Human Evolution.* Chicago and London: The University of Chicago Press.

[217.] Rilling, J. K., Glenn, A. L., Jairam, M. R., Pagnoni, G., Goldsmith, D. R., Elfenbein, H. A., & Lilienfeld, S. O. (2007). Neural Correlates of Social Cooperation and Non-cooperation as a Function of Psychopathy. *Biological Psychiatry, 61*, 1260–1271.

[218.] Roberts, D. (2005). Does the Explanatory Constraint on Practical Reasons favour Naturalism about Practical Reasons? *South African Journal of Philosophy, 24*, 97-108.

[219.] Roeser, S. (2011). *Moral Emotions and Intuitions.* Houndmills, New York: Palgrave Macmillan.

[220.] Rogers, K. (2010, September 16). *Inclusive Fitness*. Retrieved January 20, 2015, from Encyclopaedia Britannica: http://www.britannica.com/EBchecked/topic/1710067/inclusive-fitness

[221.] Rolls, E. T., Hornak, J., Wade, D., & McGrath, J. (1994). Emotion-Related Learning in Patients with Social and Emotional Changes Associated with Frontal Lobe Damage. *Journal of Neurology, Neurosurgery, and Psychiatry, 57*, 1518-1524.

[222.] Rosenberg, A. (2011). *The Atheist's Guide to Reality: Enjoying Life Without Illusions.* New York and London: W. W. Norton & Company.

[223.] Ross, D. W. (1930). *The Right and the Good.* Oxford: Oxford University Press.

[224.] Ruse, M., & Wilson, E. O. (2006). Moral Philosophy as Applied Science. In E. Sober (Ed.), *Conceptual Issues in Evolutionary Biology* (3 ed., pp. 555-574). Cambridge, Mass.: Bradford Boooks MIT Press.

[225.] Samuels, R., Stich, S., & Bishop, M. (2002). Ending the Rationality Wars: How to Make Disputes about Human Rationality Disappear. In R. Elio (Ed.), *Common Sense, Reasoning and Rationality* (pp. 236–268). New York: Oxford University Press.

[226.] Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The Neural Basis of Economic Decision-Making in the Ultimatum Game. *Science, 300*, 1755-1758.

[227.] Sarkar, S. (1998). *Genetics and Reductionism.* Cambridge: Cambridge University Press.

[228.] Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge, Mass.: Harvard University Press.

[229.] Schroeder, M. (2007). *Slaves of the Passions.* Oxford: Oxford University Press.

[230.] Schroeder, M. (2008). Having Reasons. *Philosophical studies, 139,* 57-71.

[231.] Searle, J. (2001). *Rationality in Action.* Cambridge, Mass: MIT Press.

[232.] Sesardić, N. (2005). *Making Sense of Heritability*. Cambridge: Cambridge University Press.

[233.] Shafer-Landau, R. (1999). Moral Judgement and Normative Reasons. *Analysis, 59,* 33–40.

[234.] Shafer-Landau, R. (2003). *Moral Realism: A Defence.* Oxford: Oxford University Press.

[235.] Shafer-Landau, R. (2012). Evolutionary Debunking, Moral Realism and Moral Knowledge. *Journal of Ethics and Social Philosophy, 7*(1), 1-38.

[236.] Sifferd, K., & Hirstein, W. (2013). On the Criminal Culpability of Successful and Unsuccessful Psychopaths. *Neuroethics, 6,* 129–140.

[237.] Simon, H. A. (1956). Rational Choice and the Structure of the Environment. *Psychological Review, 63,* 129-138.

[238.] Sinnott-Armstrong, W. (Ed.). (2008). *Moral Psychology* (Vols. I-III). Cambridge, MA.: MIT Press.

[239.] Sinnott-Armstrong, W. (2014). Do Psychopaths Refute Internalism? In T. Schramme (Ed.), *Being Amoral: Psychopathy and Moral Incapacity* (pp. 187-208). Cambridge MA.: MIT Press.

[240.] Skarsaune, O. K. (2011). Darwin and Moral Realism: Survival of the Iffiest. *Philosophical Studies, 152,* 229-243.

[241.] Skeem, J. L., Polaschek, D. L., Patrick, C. J., & Lilienfeld, S. O. (2012). Psychopathic Personality: Bridging the Gap Between Scientific Evidence and Public Policy. *Psychological Science in the Public Interest, 12,* 95–162.

[242.] Skorupski, J. (2010). *The Domain of Reasons.* Oxford: Oxford University Press.

[243.] Skyrms, B. (1996). *Evolution of the Social Contract.* Cambridge: Cambridge University Press.

[244.] Skyrms, B. (2010). *Signals: Evolution, Learning and Information.* Oxford: Oxford University Press.

[245.] Smith, M. (1987). The Humean Theory of Motivation. *Mind, 96*, 36-61.

[246.] Smith, M. (1989). Dispositional Theories of Value. *Proceedings of the Aristotelian Society, 63*, 89-111.

[247.] Smith, M. (1994). *The Moral Problem.* Oxford: Blackwell Publishing Ltd.

[248.] Smith, M. (2004). *Ethics and the Apriori.* Cambridge: Cambridge University Press.

[249.] Smith, M. (2009). Desires, Values, Reasons, and the Dualism of Practical Reason. (J. Cottingham, & J. Suikkanen, Eds.) *Ratio, Special Issue: Parfit's On What Matters, 22*, 98-125.

[250.] Smith, M. (2012). Naturalism, Absolutism, Relativism. In S. Nuccetelli, & G. Seay (Eds.), *Ethical Naturalism: Current Debates* (pp. 226-244). Cambridge: Cambridge University Press.

[251.] Smokrović, N. (2015). Argumentation as a Means for Extending Knowledge. *Croatian Journal of Philosophy, 15*, 223-231.

[252.] Sobel, D. (2009). Subjectivism and Idealization. *Ethics*, 336-352.

[253.] Sobel, D., & Copp, D. (2001). Against Direction of Fit Accounts of Belief and Desire. *Analysis, 61*, 44-53.

[254.] Sober, E. (1999). *Philosophy of Biology* (2nd Edition ed.). Boulder and Oxford: Westview Press.

[255.] Sober, E., & Wilson, D. S. (1998). *Unto Others The Evolution and Psychology of Unselfish Behavior.* Cambridge, Mass.: Harvard University Press.

[256.] Sperber, D., & Baumard, N. (2012). Moral Reputation: An Evolutionary and Cognitive Perspective. *Mind & Language, 27*, 495–518.

[257.] Stearns, S. C. (1989). Trade-offs in Life-history Evolution. *Functional Ecology, 3*, 259-268.

[258.] Sterelny, K. (2012). From Fitness to Utility. In S. Okasha, & K. Binmore (Eds.), *Evolution and Rationality: Decisions, Co-operation and Strategic Behaviour* (pp. 246-273). Cambridge: Cambridge University Press.

[259.] Stich, S. (1990). *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation.* Cambridge, Mass.: MIT Press.

[260.] Street, S. (2006). A Darwinian Dilemma for Realist Theories of Value. *Philosophical Studies, 127*, 109-166.

[261.] Street, S. (2008a). Constructivism About Reasons. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol. III, pp. 207-245). Oxford: Clarendon Press.

[262.] Street, S. (2008b). Reply to Copp: Naturalism, Normativity, and the Varieties of Realism Worth Worrying About. *Philosophical Issues, 18*, 207-228.

[263.] Street, S. (2009). In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters. (E. Sosa, Ed.) *Philosophical Issues (a supplement to Noûs), 19*, 273-398.

[264.] Street, S. (2016). Nothing "Really" Matters, but That's Not What Matters. In P. Singer (Ed.), *Does Anything Really Matter: Parfit on Objectivity.* Oxford: Oxford University Press.

[265.] Surridge, A. K., Osorio, D., & Mundy, N. I. (2003). Evolution and Selection of Trichromatic Vision in Primates. *TRENDS in Ecology and Evolution, 18*, 198-205.

[266.] Sušnik, M. (2015). Strong Motivational Internalism. *International Philosophical Quarterly, 55*, 165-177.

[267.] Trivers, R. L. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology, 46*, 35–57.

[268.] Turner, S. P. (2010). *Explaining the Normative.* Cambridge: Polity Press.

[269.] Tye, M. (2000). *Conciousness, Color and Content.* Cambridge, Mass.: A Bradford Book, The MIT Press.

[270.] Verbeek, B. (2007). The Authority of Norms. *American Philosophical Quarterly, 44*, 245-258.

[271.] Verbeek, B. (2008). Conventions and Moral Norms: The Legacy of Lewis. *Topoi, 27*, 73–86.

[272.] Viding, E., Blair, R. J., Moffitt, T. E., & Plomin, R. (2005). Evidence for Substantial Genetic Risk for Psychopathy in 7-year-olds. *Journal of Child Psychology and Psychiatry, 46*, 592–597.

[273.] Viding, E., Jones, A., Frick, P. J., Moffitt, T. E., & Plomin, R. (2008). Heritability of Antisocial Behaviour at Nine-years: Do Callous-unemotional Traits Matter? *Developmental Science, 11*, 17–22.

[274.] Wakefield, J. C. (2000). Spandrels, Vestigial Organs, and Such: Reply to Murphy and Woolfolk's "The harmful dysfunction analysis of mental disorder". *Philosophy, Psychiatry, and Psychology, 7*, 253-270.

[275.] Wedekind, C., & Milinski, M. (2000). Cooperation through Image Scoring in Humans. *Science, 288*, 850-852.

[276.] Wheatley, T., & Haidt, J. (2005). Hypnotically Induced Disgust Makes Moral Judgments More Severe. *Psychological Science, 16*, 780-784.

[277.] Wiggins, D. (1987). A Sensible Subjectivism. In D. Wiggins, *In Needs, Values and Truth* (pp. 185-214). Oxford: Oxford University Press.

[278.] Williams, B. (1981). Internal and External Reasons. In B. Williams, *The Moral Luck* (pp. 101-113). Cambridge: Cambridge University Press.

[279.] Williams, B. (1995). Internal Reasons and the Obscurity of Blame. In B. Williams, *Making Sense of Humanity* (pp. 35-45). Cambridge: Cambridge University Press.

[280.] Wimmer, H., & Perner, J. (1983). Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception. *Cognition, 13*, 103-128.

[281.] Wright, W. (2013). Color Constancy Reconsidered. *Acta Analytica, 28*, 435-455.

[282.] Zeier, J. D., Maxwell, J. S., & Newman, J. P. (2009). Attention Moderates the Processing of Inhibitory Information in Primary Psychopathy. *Journal of Abnormal Psychology, 118*, 554-563.

[283.] Zollman, K. J., Bergstrom, C. T., & Huttegger, S. M. (2013). Between Cheap and Costly Signals: The Evolution of Partially Honest Communication. *Proceeding of the Royal Society B, 280*.

List of tables and figures

Marko Jurjako

Curriculum Vitae

Academic Appointments

Junior Researcher on the project *Classification and explanations of antisocial personality disorder and moral and legal responsibility in the context of the Croatian mental health and care law* (CEASCRO), January 2015–January 2017.
Teaching Assistant – philosophy of mind
Fellow at the Department of Philosophy, Faculty of Humanities and Social Sciences, University of Rijeka, Sveučilišna avenija 4, 51000 Rijeka (Croatia)

Employment History

Assistant for children with special needs: Center for Autism – Rijeka, Stane Vončine 1, 51000 Rijeka (Croatia), 2013-2104.
Substitute Teacher: Osnovna škola Vežica, Kvaternikova ulica 49, 51000 Rijeka (Croatia), 2012-2013.

Education

Faculty of Humanities and Social Sciences, University of Rijeka, PhD in Philosophy, 2010 – 2016

Title: Reasons: A Naturalistic Explanation
Advisors: Nenad Smokrović and Luca Malatesti

Central European University in Budapest (Hungary), MA, Philosophy, 2009–2010

Thesis title: Moral Rationalism under Empirical Assessment

Advisor: Christophe Heintz

Faculty of Humanities and Social Sciences, University of Rijeka (Croatia), BA, Philosophy and History, 2004–2009

Thesis title: Praktično zaključivanje: razlozi i racionalnost (*Practical Reasoning: Reasons and Rationality*)
Advisor: Nenad Smokrović

Research Interests

My main interests lie in metaethics and moral psychology. In metaethics, I am mostly interested in questions about the ontological status of normative reasons and their relation to the concept of rationality. In moral psychology, I investigate the relation between rationality and the underlying neurocognitive capacities. I am interested in the ramifications of the concept of rationality when applied to moral judgments and social cognition more generally, and how the notion of rationality applies to personality disorders, most notably psychopathy. In my approaches to both metaethics and moral psychology I tend to rely heavily on empirically informed research.

Articles Published or Forthcoming

- Jurjako, M. (forthcoming). Do philosophical intuitions need calibration? *Anthropology & Philosophy.*
- Jurjako, M. and Malatesti, L. (forthcoming). The Responsibility of Psychopathic Offenders: Some Methodological Reflections. *Anthropology & Philosophy*.
- Malatesti, L. and Jurjako, M. (2016) Vrijednosti u psihijatriji i pojam mentalne bolesti (*Values in Psychiatry and the Concept of Mental Illness*). In S. Prijić-Samaržija, L. Malatesti, and E. Baccarini (eds.), *Moralni, politički i epistemološki odgovori na društvene devijacije* (*Moral, Political, and Epistemological Responses to Social Deviation*). Rijeka: Faculty of Humanities and Social Sciences in Rijeka, pp. 153-181.

- Jurjako, M. and Malatesti, L. (2016). Instrumental Rationality in Psychopathy: Implications from Learning Tasks. *Philosophical Psychology*, 26:5, pp. 717-731. **DOI**:10.1080/09515089.2016.1144876, Q1 (2014).
- Jurjako, M. (2013). Self-Deception and the Selectivity Problem, *Balkan Journal of Philosophy*, 5:2, pp. 151–162. **ISSN** 1313-888X.
- Jurjako, M. (2013). Problem intrinzično epistemičke značajnosti (*The Problem of Intrinsic Epistemic Significance*), *Prolegomena*, 10:1, pp. 83–100. **ISSN**: 18460593, Q2 (2013).
- Jurjako, M. (2011). Parfit's Challenges, *Croatian Journal of Philosophy*, 1:32, pp. 237–248. **ISSN**: 18476139, Q2 (2011).
- Jurjako, M. (2010). Uvod u teorije praktičnog uma (*Introduction to Theories of Practical Reason*), *Novi Kamov*, 35:2, pp. 37-63. **ISSN** 1333–4972.
- Jurjako, M. and Brzović, Z. (2008). Racionalnost i moralni zahtjevi (*Rationality and Moral Requirements*), *Novi Kamov*, 26:1, pp. 89–99. **ISSN** 1333-4972.