Normative Reasons from a Naturalistic Point of View

Jurjako, Marko

NORMATIVE REASONS FROM A NATURALISTIC POINT OF VIEW

MARKO JURJAKO

Marko Jurjako

Normative Reasons from a Naturalistic Point of View

University of Rijeka, Faculty of Humanities and Social Sciences
Rijeka, 2024.

Title
Normative Reasons from a Naturalistic Point of View

Author
Marko Jurjako

Publisher
University of Rijeka, Faculty of Humanities and Social Sciences
Sveučilišna avenija 4, 51000 Rijeka
www.ffri.uniri.hr

For the Publisher
Aleksandar Mijatović

Proofreading
Ema Luna Lalić
Ante Debeljuh

Cover photo
Eva Šustar

Publishing date
July 2024

ISBN 978-953-361-132-7

e-ISBN 978-953-361-125-9

This book is published with the support of the Croatian Science Foundation, University of Rijeka, and the Faculty of Humanities and Social Sciences in Rijeka.

For my Davidovich and Emmas,
may Reason be on our side.

Contents

Acknowledgments

Chapter 3 in this book was originally published as follows:

Jurjako, Marko. 2017. "Normative Reasons: Response-Dependence and the Problem of Idealization." *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action* 20 (3): 261–75. https://doi.org/10.1080/13869795.2017.1381274.

I express my gratitude to the publisher of the journal *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action* (https://www.tandfonline.com/journals/rpex20), Taylor & Francis, for granting permission to republish this material in the present book.

Special thanks go to Voin Milevski and Matej Sušnik for reading and providing comments on the whole book. Thanks also to Jeroen Hopster for reading a draft of Chapter 5.

The work on this book has been fully supported by project TIPPS funded by the Croatian Science Foundation (grant HRZZ-IP-2022-10-1788).

As always, another type of special thanks goes to the BIAS Institute and its hosts, Ivo and Sonja. Without their unwavering support and care, much of my professional work would not have been possible.

List of tables and figures

Preface

This book presents an exploration of normative reasons from a naturalistic standpoint. The book is based on my PhD thesis that was defended at the University of Rijeka in 2016. The research for this book has evolved from my enduring fascination with the concept of normative reasons—those purported facts guiding our thought and action—and the challenge of reconciling their existence within a naturalistic worldview. In the book, I present my evolving thoughts on how various aspects of normative reasons and their emergence can be understood within a naturalistic framework.

The majority of the book comprises unpublished material, with the exception of Chapter 3. This chapter, with slight changes, is based on my prize-winning essay, "Normative Reasons: Response-Dependence and the Problem of Idealization", originally published in the journal *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action* https://www.tandfonline.com/doi/full/10.1080/13869795.2017.1381274). I want to again extend my gratitude to the publisher, Taylor & Francis, for granting permission to republish this material.

I dedicate this book to my children, David and Ema. My love for them stands as compelling evidence, casting doubt on the likelihood of the truth of mind-independent realism about normative facts.

Marko Jurjako, Rijeka, Summer 2024

Introduction

The normativity of reasons and the ubiquity of the normative

Normativity pervades our thoughts and actions; for beings inherently social, such as ourselves, it appears inexorable in both individual cognition and social engagements. Onora O'Neill succinctly emphasizes its importance and pervasive presence 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. (O'Neill 1996, xi)

Normativity not only characterizes our everyday lives but is also pervasive in philosophy, the humanities more generally, and the social sciences. As is commonly asserted, for discussions of normativity (such as talk about being reasonable or unreasonable, responding to reasons, being as one ought to be, etc.) to make sense, there need 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 (Wedgwood 2007a).

In the last several decades, a discernible trend has emerged, asserting that reasons form the fundamental underpinning of normativity (M. A. Schroeder 2021). Essentially, the idea is that the concept of a normative reason serves as a foundational element upon which all other normative notions could, in some sense, find grounding (see, e.g., Parfit 2011a; Scanlon 1998; Rowland 2019; M. A. Schroeder 2021; Skorupski 2010). The notion of a normative reason is philosophically intriguing and significant precisely due to the weight it is expected to bear.

Some authors argue that the pertinent interpretation of the statement "something ought to be the case" or "something ought to be believed" is one in which the ought-claim implies that there is a decisive reason to do that thing (e.g. Parfit 2011a). Others contend that fundamental moral concepts can be elucidated through principles that reasonable individuals have a reason to accept or reject (e.g. Scanlon 1998). There are also claims suggesting that invoking reasons can elucidate how our will might be free and provide a plausible account of moral responsibility (e.g. Fischer and Ravizza 2000). Consequently, it is evident that normative reasons hold a distinctive and prominent position in contemporary discussions in ethics, metaethics, political philosophy, and the philosophy of social sciences (Gaus 2011; Logins 2022).

However, once we articulate the significance of normativity and its manifestation in terms of reasons, we must inquire into its origins and sources. This is where the puzzle emerges. As O'Neill contends, when we pose these questions, "[w]e find ourselves at sea due to the substantial disagreement about the source and authority of norms upon which we incessantly rely" (O'Neill 1996, xi).

The aim of this book is to contribute to this conversation and elucidate how the concept of reasons and their normativity might be explicated within a broadly naturalistic framework. However, before delving into this topic, we should more precisely delineate the challenges inherent in contemplating the nature of reasons from a naturalistic standpoint. Also, we should explain what we mean by normative reasons, which will be the main focus of our discussion.

Normativity, reasons, and naturalism: The problem

The challenge of providing an elucidation of the nature of normative reasons can be explained as a facet of the broader problem of explaining the phenomena of normativity as a whole within

a naturalistic worldview. Stephen Turner nicely illustrates this problem when explaining that many in the humanities and social sciences hold 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, 1–2)

The perspective that asserts a connection between every true normative claim and the existence of some normative fact raises the question of the nature of these normative facts. The pressing issue in this context is that by positing normative facts as fundamental entities, analogous to the role played by ordinary non-normative facts (such as facts about masses and forces adhering to physical laws) in providing a foundation for non-normative language, we run the risk of succumbing to the peril of introducing gaps in our worldview comparable to dualisms concerning the physical and the mental (Papineau 2002). This approach also raises concerns about invoking supernatural phenomena to explain the matters of interest. Regarding this last point, Turner, echoing John Mackie (1977), 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. (Turner 2010, 2)

In the realm of reasons, the challenge is even more immediate because the essence of reasons is believed to serve such a crucial normative function. Additionally, influential moral philosophers assert that for an entity to qualify as a reason, there must be a fact endowed with the attribute of *counting in favor of* that particular

thing for which a reason exists (see, e.g. Scanlon 1998, 18). Furthermore, some add that 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; cf. M. A. Schroeder 2007; 2021). However, (normative) reasons are meant to be those facts that genuinely impact real people with all their abilities and limitations in their actions and thoughts. Thus, it is only reasonable to expect that if reasons are real and exert influence on people, there must be a naturalistically adequate account of them. In this context, naturalistic theories confront the challenge of providing an explanation of how (normative) reasons can, in the end, be integrated into the natural world as "revealed by science" (see Harman 2000, 79).

However, before exploring the challenge of contemplating normative reasons from a naturalistic perspective, it is essential to distinguish between various senses of "reasons" applicable in these contexts.

Distinguishing between normative, motivating, and explanatory reasons

The concept of reason plays multiple roles in everyday and philosophical 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 (practical) 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 a very expensive medical treatment for his sick grandmother, and that he believed that by robbing the bank he would be able to afford it. In this example, the desire to help his grandmother and

the belief about the likely means of doing so are typically construed as providing the *reason* why Smith robbed the bank (see, e.g. Smith 1987; Davidson 2001).[1]

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 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 context of practical philosophy, explanatory reasons often take the form of motivating reasons, because they explain the actions of an agent by citing a motive for which the agent acted (see, e.g. Lenman 2009). Thus, motivating reasons explain why an action is performed by citing the reasons or considerations in light of which an agent acted.

Motivating reasons are utilized in predicting and explaining behavior and formation of the mental states of agents. Many, working in the Humean philosophical tradition, assume that motivating reasons are composed of a pair of mental state-types, such as beliefs and desires (see, e.g. Davidson 2001, essay 1; Smith 1987).[2] The theory that utilizes concepts of desires and beliefs in

1 The explanatory scheme that utilizes the notions of desire and belief in accounting for behavior or intentional action is called folk psychology (Andrews, Spaulding, and Westra 2021). 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*.

2 It is important to note that there exists an influential line of thought that dismisses a psychologistic interpretation of motivating reasons. This rejection stems from the perspective that, for the agent, motivating reasons pertain to the contents of their mental states. Consequently, some argue that motivating reasons are never strictly mental states themselves (see, e.g. Alvarez 2010; Dan-

order to explain and predict agential behavior is, in philosophical literature, often referred to as *Folk Psychology,* and in cognitive science literature as *Theory of Mind* (see, e.g. Bermúdez 2005, 33).[3]

It should be noted that nowadays it is also standard to distinguish between the more general category of explanatory reasons and strictly motivating reasons (see Alvarez 2010). For instance, if a blow to the head can cause John to believe that there are tables in front of him, we have an explanation for the formation of his belief without providing a reason in light of which this belief was formed. This is because this belief is not rationalized by some previous beliefs or perceptions that John had; rather, the new belief is just a non-rational, causal consequence of his head being struck. Moreover, the distinction between motivating and other explanatory reasons is crucial because an action may be explained by invoking a reason that is not the one motivating the agent to act. In this regard, Maria Alvarez offers a compelling illustration:

---

cy 2000). Using the example of Smith, in this perspective, motivating reasons for aiding in robbing a bank are not a belief-desire pair but rather the content or considerations that played a role in his reasoning processes—namely, the proposition that by robbing a bank, he would help his grandmother. While my use of psychologistic construals of motivating reasons may indicate a bias towards those accounts, I do not want to make a commitment to them, as my focus will primarily be on normative reasons. For an overview and discussion of different interpretations of motivating reasons, see Alvarez (2017).

3 The use of folk psychology, or theory of mind, for explaining and predicting behavior or mental states is called *mindreading* (Hutto and Ravenscroft 2021). 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 a 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 (see, e.g. Wimmer and 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 (see, e.g. Call and Tomasello 2008; Andrews 2020).

> For example, that he is jealous is a reason that explains why Othello kills Desdemona. But that is not the reason that motivates him to kill her. (…) [A]n explanation that refers to his jealousy is not a *rationalisation* of Othello's action: it doesn't explain his action by citing *his* reason. [T]he example (…) shows that not all reasons that explain by citing psychological factors, e.g., jealousy, are reasons that motivate. (Alvarez 2017, sec. 3)

Even though jealousy can explain why Othello kills Desdemona, it is not the consideration in light of which he acted. His motivating reason, in fact, was his suspicion that Desdemona was unfaithful. Thus, we observe that the reasons in light of which somebody acts can differ from the other available explanations for the same action.

Explanatory and motivating reasons are contrasted with *normative* or *justificatory* reasons (Alvarez 2017; see, also Lenman 2009). In general, normative reasons indicate how things *should* or *ought* to be, rather than describing how things currently are, or predicting their future states. It can be asserted that normative or justificatory reasons function akin to instructions that are based on the desirability or worthiness of states of affairs. As such, they are often closely linked to what is considered valuable (see Parfit 2011a, 1:38–39; Raz 1999, ch. 2) or, more generally, to those facts that determine which responses are fit to make (see, e.g. Cullity 2022).

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" (M. A. Schroeder 2007, 11; see, also Parfit 2011a, 1:31; Scanlon 1998, 17). With this terminology in mind, it is common to assert that certain facts *favor* the adoption of specific attitudes towards particular propositions. For instance, the proposition that the high concentration of iridium at the Cretaceous-Tertiary boundary *counts in favor* of the thesis that, during that period of Earth's history, an asteroid fell on Earth, leading to the extinction of dinosaurs. In the practical domain, we frequently encounter statements such as

the following: the fact that smoking cigarettes is detrimental to your health *counts in favor* of stopping smoking, or the fact that a group of people will benefit from acting cooperatively *counts in favor of* choosing to act morally.

Examining the Smith example through the lens of normative reasons highlights the crucial distinction between normative and explanatory (or motivating) reasons. In this scenario, Smith developed the belief that his grandmother was unwell based on information from a fortune-teller. However, from a normative standpoint, we can criticize Smith's formation of the belief, pointing out the unreliability of fortune-tellers as sources of information, rendering their insights insufficient reasons for belief. Additionally, when considering Smith's bank robbery, we can further criticize his actions on normative grounds, noting that he lacked a compelling reason for such behavior. From a moral standpoint, stealing and causing unnecessary harm to others is wrong, providing a basis for critique, even though we may understand Smith's reasons for carrying out the action.

The overarching idea is that it is possible to possess a reason for believing or doing something without that reason being normatively good, meaning that it does not inherently support or count in favor of that action. Conversely, one can have a normative reason for doing something without possessing a motivating reason for the action. This scenario may arise if the agent fails to recognize the reason, or even if the agent acknowledges the reason but chooses not to respond to it.

As an illustration of the first case, we can take the famous example given by Bernard Williams (1981). In this example, a person enters a bar and orders a gin and tonic. Unbeknownst to her, the bartender mistakenly pours petrol into her glass. Given her desire to consume a gin and tonic and her belief that the glass before her contains the intended drink, she proceeds to drink from the glass in front of her. While her desire and belief provide an explanation in terms of a motivating reason for why

she drank from the glass, the example highlights that, intuitively, she lacked a *normative reason* to drink it. In other words, 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 this fact strongly counts in favor of her quitting smoking. Nevertheless, due to weakness of will, the person continues to smoke when the opportunity arises, even though she recognizes that it would be better if she stopped smoking.

Now that we have differentiated between several notions of reasons, the subsequent focus in the remainder of the book will largely center on normative reasons. This is because many believe that fitting normative reasons into a naturalistic account is the most challenging task.

A naturalistic approach to normative (practical) reasons

The purpose of this book is to explore the nature of normative reasons and identify an account of them that aligns with a broadly naturalistic worldview. To streamline the book's focus, I will primarily talk about and explore the nature of normative *practical* reasons (see Chapter 1 for an outline of shared characteristics of normative reasons). Those are the reasons that pertain to actions or associated attitudes, such as desires and intentions.

By "naturalistic worldview" I refer to perspectives on the natural world presupposed in presently accepted scientific theories (for recent discussion, see De Caro 2023). Naturalism can take various forms with respect to a domain D. For example, one might argue that naturalistic principles require concepts in D to be reduced to more naturalistically acceptable concepts in another domain T, or that some concepts in D should be revised or eliminated if they do not correspond to anything in reality.

My objective is not to present a formal naturalistic reduction of concepts related to normative reasons (for such attempts, see Nuccetelli and Seay 2012). Nor do I, at the outset, feel compelled to consider normative reasons as inherently incompatible with naturalism, leading to thorough eliminativism. Instead, my stance can be characterized as methodological naturalism. While I recognize that methodological naturalism may lead to reevaluating intuitions about reasons and their ontology, its main goal in the present context is to provide an understanding of normative reasons. This involves recognizing their fundamental role in guiding agents with specific cognitive abilities and particular social and biological histories. The approach aims to achieve this by leveraging resources and insights from collaborative efforts in the social and natural sciences, addressing the complex issues of the normative.

Methodological naturalism, in a broader sense, extends beyond guiding research and philosophical theses; it also constitutes a philosophical assertion regarding the overarching relationship between philosophy and science. I regard methodological naturalism as comprising two components: one that is ontological or metaphysical, and the other that is epistemic or methodological in the narrow sense (see Papineau 2023). The ontological component pertains to the methodological maxim of grounding concepts and purported philosophical facts, such as facts about reasons and rationality, "in the world of facts as revealed by science" (Harman 2000, 79). This is just the methodological counterpart of the physicalist/naturalist claim "that reality has no place for 'supernatural' or other 'spooky' kinds of entity" (Papineau 2023).

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" (Papineau

2023). 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 2023).

I emphasize this second component of methodological naturalism because it imposes more significant constraints on philosophical theories and is related to legitimizing arguments closely connected to scientific practice (see Chapters 3 and 5, where such considerations are applied, in order to argue that naturalism favors subject-dependent views of normative reasons). In this context, it is crucial to note that one feature of methodological naturalism (construed narrowly) is the claim that the scientific method holds a certain kind of general authority. This includes the assertion that default authority should be granted to the outputs of the scientific method and its presuppositions. For instance, this constraint allows us to argue that legitimate norms of rationality are those derived from, or at least underpinned, by the relevant scientific practice or theories that employ the concept of rationality (see, e.g. Colyvan 2009; Jurjako 2022).

Moreover, methodological naturalism (narrowly construed) compels us to pose specific questions and structure our investigations in particular terms. This approach is designed to guide us in identifying a theoretical problem, determining the methods to apply for its resolution (if solvable), and assessing the feasibility of a solution. Specifically, it proves valuable to scrutinize the function of the concept of a normative reason in our discourse and inquire about its role in our mental economy. In essence, the understanding of methodological naturalism I adopt prompts us to frame the issue in the context of the problems faced by human beings (or rational agents more broadly) and how the acceptance or the introduction of the notion of a normative reason could assist in addressing these challenges (see Chapters 4 and 6, where these considerations are utilized to account for certain aspects of normative reasons).

The approach advocated by methodological naturalism stands in stark contrast to the traditional conceptual analysis approach of analytic philosophy (see, e.g. Jackson 1998; Smith 1994). Traditional conceptual analysis typically involves proposing analyses of concepts and subsequently testing them against our intuitions about the application of the concept. If counterexamples are found, the proposed analysis fails; if not, the analysis may be considered successful. This entire process is conducted *a priori*. The most notable example of this methodology is the case of the so-called Gettier problem. In his work (1963), Edmund Gettier demonstrates that our intuitive belief that the concept of knowledge can be analyzed in terms of justified true belief is mistaken, because we can conceive of cases (counterexamples) where a person has a justified true belief, yet we would not ascribe knowledge to that person.

In contrast, methodological naturalism proposes that we should not rely solely on our *a priori* intuitions about concept application. Instead, it recommends that we consider how the relevant concepts are used in successful scientific theories. Thus, part of the task of methodological naturalism is to explore how our ordinary concepts interface with scientific concepts, such as the folk-psychological concept and the scientific concept of rationality. Furthermore, this approach imposes constraints on concept application that do not solely stem from our *a priori* intuitions; it also depends on the actual usage observed in scientific theories (see Chapter 6, for such applications of methodological naturalism).

Since methodological naturalism contrasts with traditional *a priori* conceptual analysis, I should explain why I consider the former approach seriously. The answer to this question might not be completely compelling, as there is no argument that can persuade everyone to accept naturalism.[4] Some even claim that nat-

4 Some even claim that if there were reasons to accept naturalism, then natu-

uralism cannot be given any completely non-circular argument in its favor (Giere 2008). However, 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 philosophical theorizing. Instead, 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).

Nonetheless, I offer 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.

The naturalistic stance adopted in this work is rooted in an inductive inference commonly employed to advocate for the causal closure of the physical domain. This ontological proposition asserts that all physical effects can be traced back to sufficient physical causes (see the appendix in Papineau 2002). However, it is important to note that, in my discussion, I do not directly rely on the principle of the causal closure of the physical. Instead, I contend that the inductive inference drawn from the historical and current successes of empirical sciences at least justifies the attention directed toward relevant empirical sciences. This involves an effort to establish a foundation or interface of philosophical concepts with explanatory concepts derived from pertinent empirical theories. Over the past decades, adopting such an approach has proven beneficial in exploring the evolutionary, neurological, and cognitive foundations of morality (see, e.g. Kumar and Campbell 2022). Philosophers engaging with

ralism would be false, since it cannot accommodate the notion of a normative reason (Parfit 2011a). One of the main aims of the book is to show that there is a viable naturalistically friendly account of normative reasons.

scientific data have successfully formulated new perspectives and arguments, thereby advancing the discourse on traditional issues like the nature of moral judgment and its relationship to motivation. Reciprocally, these philosophical engagements have facilitated the development of scientific hypotheses and avenues of inquiry. Presently, there is no basis to assume that rigorous scientific methods of investigation and theorizing cannot be applied to domains—such as ethics—that have traditionally been regarded as primarily philosophical. Notably, scientific probing into ethics has been underway for a considerable period (see, e.g. Sinnott-Armstrong 2008). In this context, John Doris and Stephen Stich assert:

> 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 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 and Stich 2011, 112)

These considerations bring us to the second point.

The adoption of methodological naturalism in this work is driven by a pragmatic agenda central to the book. Firstly, the primary objective of this book is to explore the conception of normative reasons that emerges when scientific knowledge and relevant theories are held constant. Secondly, on a preliminary basis, an account of reasons and rationality appears more robust when it can facilitate the integration of these concepts across scientific domains and their application in various social practices. For instance, a compelling issue arises in considering how empirical data on human reasoning capacities can be leveraged to assess the rationality of individuals or to what extent they respond to reason (Samuels, Stich, and Bishop 2002; see, also, the introduction in Knauff and Spohn 2021). This discussion extends to examining whether individuals with psychopathy can be deemed

rational amoralists, a determination relevant to considerations of their liability for punishment (see, e.g. Aaltola 2014; Maibom 2018; Jefferson and Sifferd 2018). Here, the problem is that neuropsychological studies typically offer the primary evidence for adjudicating this question, and the data that these studies are based on need to be somehow integrated with our concepts of reason and rationality (Jurjako and Malatesti 2016). Thus, to engage meaningfully with such issues, it is important to include naturalistic considerations when formulating accounts of reasons and rationality that can effectively address these questions.

The necessity of relying on scientific data places us in a position where default authority is granted to both the data and the empirical theories explaining them. Additionally, empirical theories elucidating scientific data impose constraints on the conception of reason we may adopt and the norms that are expected to govern capacities associated with this concept. For instance, a natural understanding of rationality involves the capacity to adaptively respond to present and future environments in accordance with one's aims and values. This conceptualization of rationality is frequently invoked in various accounts of criminal and moral responsibility (see, e.g. Fischer and Ravizza 2000). Moreover, considering these capacities as executive functions implemented in the brain's prefrontal cortex is plausible. However, when conceptualizing reason or rationality as implemented in the brain's functions, it becomes essential to be sensitive to functions that cannot be predetermined *a priori*. One must consider what the brain is doing as implemented in the body, its role in regulating behavior and various bodily processes, and the evolutionary history shaping rationality as an executive function (see, e.g. Hirstein, Sifferd, and Fagan 2018). This external perspective on the functions of rationality and its implementation necessitates a consideration of the brain's evolutionary history and the reasons behind the evolution of rationality as an executive function. This perspective, in turn, *a posteriori* constrains, through

our scientific theories, the norms legitimately regarded as governing the proper operation of rationality (Knauff and Spohn 2021). Consequently, it extends to determining the reasons we can attribute to individuals.

These considerations compel us to take seriously methodological naturalism, which posits that *a posteriori* presuppositions derived from the sciences should both constrain and guide our arguments and the formulation of our theories. Nonetheless, even as I align with this methodological conception, I acknowledge the continued significance of the concepts that initiated our investigation. In this regard, I endorse José Bermúdez's cautionary note 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, 12–13)

Bermúdez's insight emphasizes a two-way obligation of answerability. In our exploration of normative reasons, we must be mindful of our pre-theoretical understanding while staying open to adjustments prompted by empirical findings. This dynamic interplay between foundational concepts and empirical insights is crucial in our naturalistic study. Our commitment to responsiveness ensures that our understanding of normative (practical) reasons remains grounded yet adaptable, enriched by the insights uncovered through empirical investigation.

Overview of the book

The book is structured into chapters as follows: The first chapter offers a mostly descriptive overview that introduces some of the characteristic features of normative reasons. Throughout this overview, I will rely on the explication of normative reasons as things that count in favor of something, as articulated by Derek

Parfit (2011a) and Thomas Scanlon (1998). The appeal of this explication lies in its neutrality concerning the underlying nature of reasons, as emphasized by Sharon Street (2017).

In the second chapter, building on Parfit's work (2011a), I will draw distinctions between object-based and subject-based theories of normative reasons, providing an exploration of their respective advantages and disadvantages. I contend that a naturalistic perspective on normative reasons favors subject-based theories of normative reasons. Thus, in this chapter I will offer a preliminary defense of the viability of subject-based theories of normative reasons against prominent objections.

In the third chapter, I continue with a defense of a subtype of a subject-based theory of normative reasons, namely the response-dependence theory. This defense is prompted by objections claiming that such theories lack a foundation in our common understanding of normative reasons, leading to a perceived deficiency in explanatory power concerning crucial aspects of normative reasons, such as using idealization to determine our normative reasons for action. To address these objections, I will establish an analogy between reasons and colors, illustrating how our intuitions in various domains, including normative reasons, can evolve through scientific progress. Defending a response-dependent theory of normative reasons, I will argue that certain facts become reasons due to their role in our rational responses. Additionally, I will explore how the concept of idealization can be employed in these theories to capture instances where we may be mistaken about our normative reasons.

In the fourth chapter, I will expand upon the prior discussion, highlighting additional avenues through which idealization can contribute to subject-based theories of normative reasons. Specifically, I aim to show how normative reasons, influencing people's preferences and actions, may arise from motivating reasons through interactions among minimally rational agents.

In the fifth chapter, I advocate for the endorsement of sub-

ject-based theories of reasons grounded in evolutionary and naturalistic considerations. Here, I present and defend an evolutionary debunking argument against mind-independent realism concerning normative reasons. In line with this argument, object-based theories face notable challenges, which subject-based theories of normative reasons can more effectively address. Moreover, the novel contribution of this discussion lies in my examination of insufficiently discussed arguments by Parfit (2011a) that question the effectiveness of evolutionary debunking arguments against mind-independent realist views of normative reasons. The aim of discussing and rebutting these arguments is to strengthen one form of evolutionary debunking arguments against object-based theories of normative reasons.

In the sixth chapter, I construct a subject-based theory of reasons aligning with a naturalistic perspective that aims to elucidate a crucial distinction between hypothetical and categorical normative reasons. Throughout this development, I explore various facets, including the interplay between substantive reasons, the faculty of reason, and rationality. Drawing on insights from game theory, I articulate how categorical reasons might evolve from the existing motivating or hypothetical reasons. Additionally, I explore the role of distinct forms of rationality in explaining different categories of reasons.

The chapters are meant to present a continuous flow of topics related to the question of how a naturalistically informed theory of different aspects of normative reasons could be formulated. Where one chapter opens a problem or a question that is answered in the following one. Nonetheless, despite such a structure of the book, the ensuing chapters are largely self-contained, and for the most part they can be read separately in any order the reader prefers.

Throughout the book, especially in chapters 1, 2, and 4, I will draw upon Parfit's influential work. There are at least two compelling reasons for my choice. First, Parfit (2011a; 2011b; 2017)

dedicated three volumes to the discussion of normative issues, where reasons play a special and central role, providing a robust framework for discussion of normative reasons. Second, Parfit adeptly exposes the challenges that the concept of a normative reason introduces into the naturalistic picture of the world. While advocating for his brand of normative realism, Parfit puts forth intuitively strong arguments against naturalism regarding normative reasons (see, also the papers in Singer 2017; Nuccetelli and Seay 2012). Thus, beyond providing a framework for discussing reasons, Parfit serves as an intellectual opponent of great importance in modern philosophy, whose arguments need to be addressed to demonstrate that normative reasons can be incorporated into a naturalistic worldview (Edmonds 2023).

1 Features of normative reasons

1.1 Introduction

The aim of this chapter is to clarify the concept of a normative reason and outline the preliminary perspectives on its nature. To do this, I examine its structural characteristics, including its relations to other normatively relevant concepts, such as rationality, deliberation, and advice.

1.2 Practical and theoretical normative reasons

Within the realm of normative reasons, a standard distinction exists between theoretical or epistemic reasons and practical reasons or reasons for action. Broadly speaking, this distinction can be framed in folk-psychological terms. By utilizing folk psychology, we can, among other things, explain and predict the behavior of intentional agents by ascribing mental states to them (for discussion, see Andrews, Spaulding, and Westra 2021). These mental states are commonly categorized into two broad groups, broadly referred to as cognitive states and motivational states. Cognitive states encompass beliefs, suppositions, assumptions, plausibility judgments, etc., while motivational states encompass desires, intentions, emotions, preferences, and the like. For the sake of simplicity in our discussion, cognitive states are often collectively referred to as "belief", and motivational states are subsumed under the term "desire" (see, e.g. Smith 1987). This terminology has led to the common characterization of folk psychology as belief–desire psychology.

By employing this classification, the differentiation between epistemic and practical reasons can be articulated in relation to reasons that support distinct types of mental states. Theoretical reasons, in this context, are grounds to believe something or to adopt a belief concerning a particular state of affairs. For instance, the discovery of iridium in the Cretaceous-Tertiary layer

of the Earth's crust provided scientists with a reason, given their other theoretical beliefs, to believe that dinosaurs became extinct due to the impact of an asteroid (for an overview of epistemic reasons, see Introduction in Reisner and Steglich-Petersen 2011). In contrast to epistemic reasons, practical reasons contribute in favor of actions, desires, and intentions; more broadly, they pertain to the type of motivation one should have. For instance, it is generally accepted that if a person is in pain, we have a reason to assist them in alleviating that pain.

However, the broad distinction between epistemic and practical reasons is perhaps more intuitively understood in terms of rational requirements that apply to motivational and epistemic states. Gilbert Harman (2004) provides a good example of different requirements that apply to intentions and beliefs. Suppose 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—each demanding a similar amount of effort, equal in distance, equally boring, safe, and so on. Given these comparable features of the routes, it is rational for me to choose one arbitrarily. Since the two routes are similar in all relevant respects, it is entirely rational to decide, for example, by flipping a coin. However, in the epistemic case, an analogous situation would not warrant the arbitrary adoption of a belief. If I am in a situation where 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 adopting various attitudes help illustrate the distinction between practical and epistemic reasons. Practical reasons appear to be considerations that meet the rational demands of practical attitudes, like intentions in our example. On the other hand, epistemic reasons are distinct; they are considerations that fulfill the rational demands

applicable to beliefs, for instance. Thus, intuitively, we observe a close connection between the facts providing reasons of different types and the rational requirements governing the diverse attitudes for which we seek these reasons.

In addition to the overarching division between epistemic and practical reasons, we can further delineate subdivisions within normative reasons. These may include aesthetic reasons, reasons of etiquette, moral reasons, legal reasons, and so forth. In this context, one might inquire about the relationship between, for instance, epistemic and practical reasons and whether one type can be reduced to the other. However, for our current purposes, this issue is not paramount. In what follows, I will focus on the general concept of normative reasons to elucidate certain structural features that will then serve as a foundation for further discussions about the nature of normative reasons.

1.3 Commonly observed features of normative reasons

John Skorupski suggests that certain features of reasons can be discerned from the following general form of the reason-relation:

> Set of facts r_i is at time t a reason of degree of strength d for X to ψ. (Skorupski 2010, 37)

Here, r_i represents facts that favor something (the ground or basis of the reason relation), t signifies time, d denotes the strength of the reason or reasons in question, X represents an agent to whom the reason relation applies, and ψ stands for the thing the grounds are reasons for—whether it is a belief, action, desire, or some other attitude.[5] Usually, in discussions about normative reasons, reference to time is omitted, so in my discussion, I will also not include the temporal dimension of reasons.

5 Just a side note, Skorupski (2010, 35–36) identifies three fundamental types of reason-relations: reasons for belief, action, and feeling, which he terms epistemic, practical, and evaluative reasons, respectively.

For example, intuitively, we can assert that the fact that this glass contains petrol is a strong reason for Mary *not* to drink from it. Likewise, we can argue that the fact that Smith has seen the fossil records of different organisms is a reason for him to believe that evolution occurred.

1.3.1 Reason is a relation

The formal structure of propositions employing the concept of reasons reveals that reason is a *relation* between facts and attitudes (see, also Broome 2021).[6] It also illustrates that the relata of the reason-relation encompass a basis or ground constituted by certain facts and an attitude for which those grounds count in favor. For instance, the fact that the clouds are grey serves as a ground for the reason-relation supporting the belief that it will rain—the other relata of the reason-relation. Moreover, the greyness of the clouds supports the belief to a certain degree, given that the connection between the grey clouds and the likelihood of rain is probabilistic.

It is crucial to emphasize that reason claims are relational, as reasons are often associated with facts that constitute the ground or basis of the reason-relation. While there is little harm in linking reasons to these facts, stressing that reasons are relations, captured by the phrase "counts in favor of", helps us avoid potential conundrums. For instance, it resolves questions about how reasons can be ordinary descriptive facts, such as the fact that I am in pain, while simultaneously being normative in the sense of indicating that something needs to be done. The solution lies in recognizing that the fact that I am in pain cannot be

6 Earlier, I mentioned that practical reasons could serve as reasons for action. However, actions are not attitudes. To reconcile this apparent gap, following Thomas Scanlon (1998) we can posit that reasons for action are mediated by reasons for intention. Given that intentional action typically arises from an intention to perform an action, we can thus maintain the connection between reasons and attitudes.

equated with the fact that I have a reason to change my situation. It is philosophically more precise to state that the fact that I am in pain counts in favor of altering the current situation in some way.[7]

1.3.2 Pro tanto and prima facie reasons

The degrees of support that reasons bring with them indicate their *pro tanto*[8] nature (see, e.g. Broome 2013). *Pro tanto* reasons are those that genuinely support ψ-ing, but their degree of endorsement for ψ-ing may not be decisive; it could be outweighed by other, stronger reasons. For instance, the fact that the glass contains petrol is a reason not to drink from it. However, consider a scenario where Mary's drinking petrol could save a person's life. Suppose malicious individuals threaten to kill Mary's friend unless she drinks the petrol from the glass. In this case, the fact that drinking the petrol could save a life becomes a reason for Mary to drink it or at least to form an intention to do so. Nonetheless, the fact that drinking petrol could make Mary sick remains a reason not to drink it, albeit a reason outweighed by the stronger reason to save a friend's life. James Lenman (2009) offers another example. We can assume that the fact that smoking brings Mary pleasure provides *a pro tanto reason* for her to smoke. However, even though there might be something speaking in favor of Mary's

7 For further discussion, see Skorupski (2010).

8 Different authors articulate the notion that reasons can be *pro tanto* in various ways. Historically, they were often referred to as *prima facie* following David Ross' (1930) distinction between *prima facie* and absolute duties. However, the term "prima facie" implies that what initially seemed like a reason might not be a reason at all. To revisit Williams' example, the fact that Mary ordered a gin and tonic is a *prima facie* reason to drink the beverage given to her by the bartender. Yet, the revelation that the glass contains petrol negates the *prima facie* reason. In other words, if Mary discovers the contents are actually petrol, she will realize that she does not have a reason to drink it. The term "pro tanto", however, allows even outweighed reasons to maintain their status as reasons in favor of something. For an example of *pro tanto* reasons, refer to the main text. Dancy (2004), for instance, employs the term "contributory reason", for what other philosophers referred to as a *pro tanto* reason.

smoking, we might still conclude that, *all things considered*, Mary should not smoke.

Reasons can also be *prima facie*. Unlike *pro tanto* reasons, *prima facie* reasons can be defeated, not merely outweighed. To illustrate, let us once again revisit Williams' petrol example. The fact that Mary ordered a gin and tonic is a *prima facie* reason for her to drink the beverage given to her by the bartender. However, the fact that the glass contains petrol cancels out or defeats the *prima facie* reason. In other words, if Mary were to discover that the glass actually contains petrol, she would realize that she does not, in fact, have a reason to drink the beverage. Thus, when Mary recognizes the presence of a defeater for her reason to drink the beverage, that reason ceases to count in favor of consuming it.

In this context, it is helpful to employ a distinction introduced by John Pollock (1987, 485) 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 concept of a rebutting defeater. The fact that Mary ordered gin and tonic supports the reason to drink the contents of the glass given by the bartender. However, the fact that the glass contains petrol supports the reason not to drink the contents, and consequently, the latter reason defeats the former. Undercutting defeaters are those that undermine the connection between the reason and what the reason supports (the conclusion). For instance, we can assert that the fact that it appears to us that Smith is in pain is a reason to help him. However, the fact that we are in a theater, and Smith is an actor, undermines the conclusion that we should help him.

1.3.3 Reasons and deliberation

One of the pivotal roles that reasons play in our mental lives is evident in the deliberation about what to do or what to believe (see Enoch 2011, ch. 3). When faced with choices between different

possible actions or when deliberating about what to believe, conflicting reasons pertaining to the issue can arise. In the context of rational decision-making or endorsing a belief, the choice depends on the strength, weight, or, one might say, the force of reasons (see Parfit 2011a, 1:32). Reasons can be amalgamated so that, on one hand, there is a compelling reason to opt for a particular course of action, while, on the other hand, there are several individually weaker reasons that, when combined, become stronger than the first. Parfit provides an intuitive compelling 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. (Parfit 2011a, 1:32)

Parfit also introduces the concepts of decisive and sufficient reasons. A decisive reason to act exists when "our reasons to act in some way are stronger than our reasons to act in any of the other possible ways". Additionally, Parfit explains that acting in accordance with the decisive reason "is what we have most reason to do" (Parfit 2011a, 1:32).

However, the concept of a sufficient reason is introduced because intuitively, there will be situations in which there is no decisive reason to do any particular thing, yet there is still enough reason to act in more than one way. As Parfit notes:

> 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. (Parfit 2011a, 1:33)

Reasons and their role in deliberation can exist at different levels. While we might have a first-order reason to do ϕ, there could also be a second-order reason to disregard the first-order reason in a

particular situation.[9] Joseph Raz defines second-order reasons as "reason[s] to act on or refrain from acting on a reason" (Raz 1975, 34).

Scanlon (1998, 51) gives the following example that can be used to illustrate the distinction: when engaged in a game of tennis, the decision to play competitively or not arises. Assuming the choice is to play competitively, the fact that a particular shot represents the best strategy for winning a point becomes a sufficient reason to execute it. In this context, there is no need to weigh this reason against the possibility of causing discomfort to the opponent or hurting their feelings as a result of the competitive play. Although there might be reasons to care about the opponent's feelings, they are deemed irrelevant in the tennis match setting. Thus, in this example, we may have a first-order reason to consider our opponent's feelings. However, during the tennis match, these reasons are disregarded due to the presence of second-order reasons, particularly those based on the decision to play competitively.

The normative element of reasons becomes prominent when we incorporate the concept of an "ought" into discussions about reasons as facts that support a particular course of action. The link between what ought to be the case and the favoring relation becomes evident when contemplating what to do or believe and subsequently forming a judgment about having a decisive reason for those actions or beliefs. In such instances, it is natural to assert that what we have a decisive reason to do is what we should or ought to do. According to Parfit (2011a), the crucial sense of "ought" in the context of normative reasons implies the presence of a decisive reason 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

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

reason should in some sense explain or justify the ought-claim (Logins 2022).[10] 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 considered, he has a decisive or at least sufficient reason to help Smith.

Other forms of deliberation can occur 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.

1.3.4 Reasons, rationality, and advice

Other prevalent perspectives associated with reasons include the idea that normative reasons are those considerations that could

10 See Broome (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. Other prominent accounts construe normative reasons as answers to normative questions. For recent discussion, see Logins (2022).

be offered as advice regarding what actions to take or what beliefs to adopt (see Smith 1994; 2004; Manne 2014; cf. Arkonovich 2011). The fact that the glass contains petrol serves as a reason for Mary not to drink from it, and because it functions as a reason for her not to drink, someone in a superior epistemic position to Mary could proffer this fact as a piece of advice, advising her not to drink from the glass.

This view is based on the common idea that an agent does not necessarily have to be aware of all the normative reasons that are applicable to her in a specific situation. Building on this correlation between advice and reasons, Michael Smith (1994; 2004) has formulated a theory of normative reasons, suggesting that an agent has a reason to *Φ* if her rational self would desire that she *Φ*. In other words, if her rational counterpart, possessing comprehensive knowledge of her and her circumstances, would advise her to *Φ*, then she has a reason to do so.

We can draw another distinction in terms of the relationship between rational advice-giving and awareness of reasons—specifically, the distinction between subjective and objective reasons (see, also M. Schroeder 2008). Subjective reasons refer to those reasons an agent believes she has, while objective reasons are those that genuinely apply to her. When one is not aware of a reason, it is possible to act against that reason, doing something that one ought not to do from the perspective of that reason. In such cases, a person may appear to act against a reason without being irrational, given the beliefs in the light of which she acts. The petrol example illustrates this point effectively. Mary has a reason not to drink the contents of the glass; however, if she actually drinks the petrol, she would still be rational, at least in a minimal sense, as she would be acting according to her justified belief that the glass contains what she ordered. In a derivative sense, her action would be justified from her own subjective perspective, acting for a reason that she believes to obtain.

In this regard, subjective reasons could be associated or identified with motivating reasons. Several philosophers construe motivating reasons as normative reasons in light of which an agent acts (see, e.g. Alvarez 2010; Dancy 2000; for discussion, see Alvarez 2017). In this view, subjective reasons can be understood as motivating reasons that, through their connection with an agent's rationality, explain their actions. Objective reasons, on the other hand, could be understood as capturing normative reasons more directly. While this view successfully captures the content of some motivating reasons, not everyone agrees that the content of motivating reasons will necessarily correspond to the content of a subjective reason, understood as a belief that there is a normative reason to perform some action (for discussion, see Mantel 2018). For instance, a naturalist perspective in philosophy should remain open to the possibility that eliminativism about normative reasons is true—namely, that there are no facts that count in favor of something (Olson 2014). A person might hold this view and, consequently, think that there are no normative reasons (for discussion, see Streumer 2023; Taccolini 2024). However, she could still intentionally act based on some motivating reason that rationalizes her action. In this context, the reason in light of which this person acts will not necessarily be a belief in a normative reason that favors the action. Since this perspective does not immediately reveal a contradiction or incoherence, it is advisable not to commit, at this point, to equating motivating reasons with subjective reasons.

Another thing that could be noticed from the distinction between subjective and objective reasons is that the concept of rationality may be linked to a more subjective understanding of a reason. This is because our judgments of rationality of an action or belief often rely on the contents of a person's desires and beliefs. This view contrasts with the notion of an objective reason, which is more extensional, meaning that what there is a reason to want and believe depends on facts rather than strictly on the

mental state of the agent. We can employ Parfit's example to illustrate this 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, 1:34)

The example highlights the potential disparity between our beliefs about what is rational and what we, when adopting a third-person perspective, recognize as a reason to act and would advise ourselves to do. In this regard, the examples of the gin and tonic and the disturbed snake seem to show that our intuitions regarding rationality and our reasons for action may diverge. Drawing from these observations, some authors infer that the intuitive notion positing rationality as a response to reasons might be erroneous, or at the very least, not as unequivocal. This is grounded in the idea that one can exhibit rational behavior even in the absence of adhering to an externally determined reason (Broome 2013). According to this view, rationality can be conceptualized as a set of requirements that govern the appropriate combination of mental states, irrespective of the reasons supporting those attitudes. For instance, within this framework, rationality may necessitate the intention to escape if one desires to preserve one's life and believes that doing so is contingent on fleeing. Failing to generate the intention to run away, despite the existing mental states, would appear to result in an irrational combination of mental states, independent of external circumstances. In this regard, reasons and rationality may diverge, as reason could dictate remaining in place (unbeknownst to the individual), while rationality

would mandate fleeing based on their current attitudes.[11] Thus, some argue that rationality diverges from reasons, as reasons are grounded in external facts, whereas rationality supervenes on mental states, irrespective of which facts constitute reasons for particular actions (see, e.g. Broome 2013).

Proposing a conceptual distinction between rationality and reasons leads to problems. Typically, we perceive rationality as normative, implying that we ought to adhere to the rules of rationality, and deviating from them suggests that something is wrong. However, if rationality is exclusively tied to fulfilling coherence criteria in the formation of beliefs and other mental states and is not inherently linked to responding to reasons, the question arises: why should we prioritize rationality, and what reason do we have for being rational? (Kolodny 2005) If we maintain that rationality merely requires coherence among our attitudes, it becomes challenging to provide a principled response to the latter question—namely, to articulate what would be inherently wrong with lapses in rationality (for discussion, see Lord 2018; Kiesewetter 2017).

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

11 For advocates of the perspective positing that rational requirements are wide in scope, rationality demands either forming an intention to retreat or adjusting one's attitudes to restore coherence among mental states. This can be intuitively illustrated through theoretical reasoning involving *modus ponens*. Assuming *modus ponens* inferences represent a facet of rationality, proponents of the wide-scope view posit that rationality entails the following: suppose agent A believes that p, believes that if p then q, and believes not-q. Given that this combination of beliefs is inconsistent, rationality dictates either ceasing to believe not-q or revising the belief in p, as either adjustment would reinstate consistency. Formally, expressing *modus ponens* as a rational requirement can be articulated as follows using an ought operator with a wide scope: Ought (if you believe that p ∧ you believe that (p → q)), then believe that q. It is essential to note that this requirement does not compel one to believe q given the other beliefs; rather, it is satisfied by either not believing that the conjunction (p ∧ (p → q)) is true or simply believing that q is true (for discussion, see Broome 1999).

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 moot, 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. 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.[12]

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 (see, e.g. Parfit 2011a, 1:111; M. A. Schroeder 2007, 14; for discussion, see Sylvan 2015). 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 rea-

12 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 (see Johnston 1989, 158).

sons when background conditions are normal. We would minimally include having relevant true beliefs among background conditions. Thus, if background conditions were normal, exercising our rational capacities would tend to lead us to what there are reasons to do. However, if background conditions were 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).

1.4 Summary: The structure of reasons

The information provided on the structural characteristics of reasons can be summarized in **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 deliberation and reasoning.

Table 1

Having outlined the concept of a normative reason, the next chapter will explore the nature of the "counting in favor of" relation, examining the truth conditions of such claims and assessing their potential integration into "the world of facts as revealed by science" (Harman 2000, 79).

2 Ontological accounts of normative reasons

2.1 Introduction

The goal of this chapter is to introduce two categories of theories that seek to explain the nature of normative reasons. In particular, in the rest of the book, I will mostly be concerned with normative practical reasons, those that pertain to action. They can be differentiated by answers to the question: what makes claims that some fact is a normative reason to Φ,[13] or that a fact counts in favor of Φ-ing true or false? At the most general ontological level, there are two 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 facts about normative reasons are mind- or subject-dependent. The question under consideration can be put in terms of Euthyphro's dilemma. Is there a reason to believe, desire, 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.? (see, e.g. Enoch 2005, 763–64).

Following Parfit's (2011a) discussion, I will categorize normative reasons into two accounts: object- or value-based theories and subject-based theories. Object-based theories align with the idea that what confers normativity to some facts is mind or judgment independent, while subject-based theories are more congruent with the view that normative reasons are in some way mind or subject dependent. As we will see, subject-based theories align more congruently with a naturalistic perspective that emphasizes the interconnectedness of human experience and cogni-

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

tion with the world.

The overarching goal of this chapter is twofold. First, to underscore the primary issues with object-based theories. Second, to illuminate the challenges confronting subject-based accounts while exploring potential solutions. Concerning the latter, I will articulate the key challenges posed by Parfit to subject-based theories and endeavor to present plausible responses from the perspective of a subject-based theorist.

2.2 Object-based theories of 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, 1:45). 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" (Parfit 2011a, 1:45). Furthermore, Parfit explains why object-based theories can be called *value-based* theories:

> 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, 1:45)

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 (for discussions of similar views, see Scanlon 1998; 2014; Alvarez 2010; Enoch 2011; Rowland 2019). Furthermore, we can add that negative reasons or reasons for avoiding something are provided by facts that make acts or outcomes bad in some way.

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 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 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 understanding 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 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 2011b, 2:486). According to this view, truths about reasons are *necessary,* and their status is often

compared to mathematical and logical truths (Parfit 2011a, 1:129; 2011b, 2:307, 326, 489, 643, 746). 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 logical and mathematical truths, we can discover some normative truths merely by thinking about them. (Parfit 2011b, 2:489–90)

Since it is normally thought that mathematical and logical truths can be *discovered* through mathematical reasoning and reflection,[14] 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.

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 or functions is to motivate, direct, or govern actions and beliefs (Korsgaard 1986; Smith 1994). 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 to 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.

14 Or in Parfit's words: "We often *can* discover logical or mathematical truths merely by thinking about them" (Parfit 2011b, 2:489).

This seems to be a big and an important disanalogy between the necessity of mathematical truths and the necessity of truths about normative reasons (for further discussion of this issue, see Clarke-Doane 2020).

Nevertheless, Parfit (see, e.g. his 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 normative 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 (see, also Dreier 2015):

> 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, 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.[15]

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.

2.3 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:

> [O]ur 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, 1:45)

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, concerns, or generally what people care about. Since subjectivist theories are based in some way on an agent's desires, this group of theories can also be called de-

15 From a naturalistic perspective, Korsgaard's criticism might be further developed in different directions. In chapter 5, I will defend an evolutionary-based version of this criticism.

sire-based theories (for proponents of such views, see, e.g. Goldman 2009; M. A. Schroeder 2007; Smith 1994; 2004; 2013; Williams 1981; 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 '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 intrinsic (i.e. non-instrumental) desires, then we could get different predictions about what our reasons would be, rather 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, thus losing 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 (idealizing) 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. Perhaps the most general characterization that Parfit provides would be the following:

> 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 [non-instrumental] desires or aims, or is what, after ideal deliberation, we would choose to do. (Parfit 2011a, 1:64)

To further illustrate an account of subject-based theory of (practical) normative reasons, I will rely on Bernard Williams' influ-

ential theory of internal reasons (Williams 1981; 1995; for more recent defenses of normative reasons internalism, see, e.g. Markovits 2014, ch. 3; Manne 2014; Asarnow 2019).

2.4 Internalism, subject-based theories, and the normativity of reasons

In his seminal paper 'Internal and External Reasons', Williams ask us to consider the following sentences: "There is a reason for A to Φ" and "A has a reason to Φ" (1981, 101). We may wonder about the truth-conditions of these sentences. On object-based theories of normative 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, 101).[16] In his later work, Williams dropped the sufficiency condition and gave the following fuller explication.[17] 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, 35)[18]

In contrast, externalist theories, in line with object- or value-based theories about reasons, would claim that whether A has a reason

16 Williams (1981, 102) calls his interpretation *the sub-Humean model* because it is in the general spirit of 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; cf. Schafer 2015a).
17 This formulation is also present in his (1981) paper. Nevertheless, Williams (1995, 35) continues to claim that his formulation of the truth-conditions for reason-statements also provide a sufficient condition.
18 Briefly, Williams (1981, 102, 105) calls an agent's motivational set S and members of that set desires, 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.

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, it is important to address an 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' account of reasons then, in effect, we eliminate their normativity. Therefore, according to Parfit, that account of reasons cannot provide a proper analysis of normative 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, 123)

Parfit objects that (C), if it is a statement about normative reasons, cannot be just a restatement 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 (see Parfit 1997, 126). In contrast, reason-statements are supposed to tell us what we *should* do or rationally ought to desire.

However, this objection is not persuasive. As Parfit (1997, 125) himself recognizes, Williams provides truth-conditions for statements about reasons in terms of *rational* deliberation or *sound* deliberative routes (see, also Roberts 2005, 101). In this re-

gard, (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 causal 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 causal, 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, 1:32).[19]

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

19 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, 17).

saying that "we must want, and do, what we know that we have most reason to want and do" (Parfit 1997, 116), which "could be true even if, […] no amount of informed deliberation would in fact motivate [an agent]" (Parfit 1997, 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, 2:452).

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' type of internalism is not about *normative reasons*.[20] As Sharon Street (2017) points out, even if we grant that the notion of a normative reason cannot be reduced to a psychological notion of a motivating reason, or any other non-normative notion, it still does not follow that understanding the notion of a normative reason entails the falsity of internalism about normative reasons.

This is because we can determine our normative reasons by pointing "to a certain type of conscious experience with which we're all intimately familiar", whose "intrinsic character (...) cannot accurately be captured or described except by invoking normative language" (Street 2017, 126). To render the analogy more vivid, Street compares this with our color experience, where, "for example, the intrinsic character of the experience of redness cannot accurately be described except by invoking color language (...)" (Street 2017, 126). However, Street further notes that this does not disable us from identifying these types of experiences by referring to common circumstances that give rise to those ex-

20 For a similar point, see also the appendix on Williams in Scanlon (1998).

periences. For instance, the experience of redness can be referred to by indicating that this is what it is like to see a ripened strawberry. Similarly, we can identify normative experiences associated with reasons, such as those that there is something that counts in favor of having some attitude or performing an action, by indicating that this is the type of experience that we tend to have "when a car suddenly swerves toward us on the highway, or when we see a child in pain" (Street 2017, 126).

Here, the important point is 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" (1981, 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 a 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, 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, excluding causal means-ends reasoning, of how a person might come to the conclusion that she has a reason to act in some way, include considering the order in which to satisfy desires or preferences, reconciling conflicts between them, exploring ways to combine desires for more comprehensive satisfaction, and deciding on a more conceptual basis what types of things one likes and how they can be implemented in specific life circumstances (see Williams 1981, 104).

Whether some other more substantive norms or patterns of practical reasoning necessarily belong to an agent's motivational set is a matter of dispute (see, e.g. 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 a person is advised that there is a reason for treating his wife kindlier. Despite the suggestion, the individual, known for being resistant, responds with a blunt rejection, stating that he genuinely does not care. Various attempts are made to provide different perspectives and involve this person in the matter, but it becomes evident that based on his current motivations he does not have any reason for to exhibit increased kindness towards his wife in the given circumstances (see Williams 1995, 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" (1995, 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 take other people for granted without a sense of regret (or use them in more devious ways), seem to be ubiquitous in our society (Hare 1993). Williams contends that individuals of this kind, assuming their reasoning capacities are otherwise rational, would lack any inherent reason to adhere to

moral prescriptions—such as being kind to one's spouse, avoiding harm to others, or apologizing for wrongdoing—that are typically accepted without question.

Even though Williams is skeptical of this, some subject-based theorists of reasons would claim that prudential and moral norms necessarily belong to the motivational sets of every rational agent (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 a 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.[21] What is important is that subject-based theorists, according to my construal, endorse some kind of dispositionalist or even constructivist account of reasons (see 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 idea is that the relation *counting in favor of* can be explained by examining the interaction between the rational agent's structure and the environment she is situated in. In essence, the subjectivist perspective asserts that things hold value, or provide reasons, based on their alignment with our desires and fundamental concerns, primarily determined by

21 From the discussion in chapter 5, 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.

what we currently value or would value under specific conditions (see, also Goldman 2009)

2.5 Comparing object-based and subject-based theories of normative reasons

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. 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. Instead of asserting that desires serve as reasons on subject-based theories, these theories can also acknowledge that reasons are, in fact, facts or states of affairs that can become the 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 object-based accounts they are irreducible normative facts, while on subjectivists accounts reasons are based on the "valuing subjects" (Goldman 2009, 28).

Besides the ontological difference in the latter sense, Parfit (2011a, 1:46–47) claims that subject- and object-based theories can be differentiated by what those theories imply we have or do not have a reason to do, or to want. According to Parfit, there

are principled and deep disagreements between the implications of the two theories. In light of this claim, he challenges subject-based theories by asserting that they yield implausible consequences concerning both the reasons we possess, and the reasons we believe we have. Subject-based theorists, however, hold differing views on this matter. Some argue that subject-based theories can accommodate the intuitions endorsed by object-based theorists, while others suggest revising our intuitions. To assess this concern, I will scrutinize the argument presented by Parfit, which questions the plausibility of subject-based accounts.

2.6 Subject-based theories of normative reasons and their implications

2.6.1 The agony argument

Parfit initiates his objection with the so-called agony argument, presuming that we inherently possess decisive or, at the very least, sufficient reasons to strive to avoid all future agony. This assumption forms the basis of the argument, which is as follows:

> Suppose that (…) 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, 1:73–74)

The idea is that according to subjectivist (or subject-based) 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.[22] Therefore, according to Parfit (2011a, 1:76), subjectivist theories are false.

22 Parfit (2011a, ch. 4) offers other similar examples such as the future Tuesday indifference example. 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.

Countering this argument, one could contend that even if it is logically possible for some agents to lack the desire to avoid all future agony after ideal deliberation, this might not be true for actual rational agents. Nevertheless, Parfit offers a rejoinder 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, 1:76–77)

Here Parfit assumes that certain assertions about reasons must be true across all possible worlds similar to ours for them to be true in the actual world. Parfit's paradigmatic example appears to be the desire to avoid all future agony.

Nevertheless, this appears to reflect Parfit's bias, influenced by his perspectives or intuitions regarding the nature of reasons and the theories that explain them. It appears that a "subjectivist" could offer at least two potential responses. A subjectivist might consistently argue that claims about reasons are contingent, in the sense that they depend on our rational dispositions to consider certain facts as reasons. However, it is unnecessary for subjectivists to assert that there will always be a fact or a state of affairs universally recognized as a reason for something in all possible worlds for all rational agents. This notion can be likened to Williams' contention that there may not always be a definitive answer to the question of what a person has a reason to do, because it will not always be clear what would be a conclusion of rational deliberation starting from some contingent set of desires, projects, and values:

> 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[23] 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, 110)

As evident from the quote, Williams holds the view that the reasons we possess cannot always be ascertained on *a priori* grounds. He argues that an account capable of capturing and explaining this characteristic of normative reasons is, in fact, superior to alternative explanations.

At this juncture, Parfit might counter that possessing both a reason and an awareness of this reason to desire the avoidance of all future agony is so fundamental that it cannot rely on contingent opportunities and possibilities for practical reasoning. It may appear that Parfit's argument is valid, suggesting that having a reason to desire the avoidance of future agony cannot be desire-based if we accept that it is logically possible for an agent to lack the desire to avoid future agony after ideally rational deliberation. 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 ideally 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 (2009) 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.) pertain to demonstrate. In this

23 In our current terminology, we might refer to this as a subject-based conception of reasons.

regard, 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 we could 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 (for Parfit's response, see his 2017, 3:259–63).

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" (2011a, 1:77). 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 (see Smith 2009).

In this regard, 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, 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, 1:76)

However, from a subjectivist standpoint, there are two plausible ways to support a subject-based theory of reasons. One approach is to scrutinize the rationality of the individuals in the example. Depressed people, in particular, are frequently considered to be paradigmatic examples of individuals whose rationality is somewhat impaired (see, e.g. Goldman 2007). Therefore, the strength of Parfit's argument in the above quote is not entirely clear.

Second, it might be questioned whether it is really the case that these people really do not have 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 (see, e.g. 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 the upcoming discussion, I will examine another argument put forth by Parfit that, in my view, holds greater significance. This argument contends that subject-based theories lack coherence and should therefore be dismissed. In the subsequent section, I will outline this argument, and endeavor to show that it

does not undermine subject-based theories of practical reasons.

2.6.2 The incoherence argument

The incoherence argument[24] includes a statement that aptly characterizes a variety of subject-based theories of reasons, along with a second statement articulating the conditions necessary for the truth of the first statement. According to Parfit, the crux of the argument is that a subject-based theorist cannot acknowledge the veracity of the second statement, rendering their position incoherent. Thus, the initial statement (M) in Parfit's argument is as follows:

> (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, 1:93)

Parfit introduces an additional condition, seemingly innocuous, yet reasonable from an epistemological standpoint. This condition is expressed as statement (N):

> (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, 1:93)

Parfit contends that (M) and (N) could only be true if statement (O) is also true:

> (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, 1:93)

Parfit (2011a, 1:93) illustrates this with the example of juries. He reasonably suggests that juries should reasonably consider relevant facts that provide them with reasons to believe in the guilt

24 Parfit's incoherence argument should not be confused with Michael Smith's *incoherence argument,* as labeled by Shafer-Landau (1999).

or innocence of the accused, based on which they should form a final verdict. Similarly, he argues that, especially in important life situations, individuals should strive to identify and rationally contemplate the possible outcomes of their actions when deciding which results to pursue.

Parfit argues that a subjectivist endorsing (M) and (N) cannot coherently accept (O). This is because (O) aligns precisely with what object-based theories embrace and what subject-based theories (should) reject. Parfit posits that, given the presupposition of (O) in (N), subject-based theories cannot accommodate (N) (Parfit 2011a, 1:94). Furthermore, in Parfit's view, subjectivists cannot endorse 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, 1:94)

Parfit's incoherence argument asserts that subjectivists, by accepting premises (M) and (N) that inherently rely on (O), find their position contradictory because they reject (O).

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. To use Williams' (1981; 1995) 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) carries an epistemological or methodological character, guiding how one should behave and think when making significant decisions. The purpose of statement (O) is to elucidate why something akin to (N) is employed to determine our reasons for action. Furthermore, in (O), Parfit appears to assume that the rationale for utilizing (N) in decision-making must hinge on the presence of object-given reasons arising from intrinsic features of specific 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 (see, also Parfit 2017, 3:262–63). In other words, a subject-based theorist 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 that (O) does not have to be true in order for (N) to be true. It is enough that something like (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 at hand is whether (O') can account for (M). I posit that (O') can indeed explain (M), albeit at the cost of rendering (M) an analytical statement. If (O') is valid, the reason we believe that our post-ideal deliberation desires align with what we have the most reason to do is conceptually linked with what we would desire after undergoing an ideal deliberative process.

The question of whether this poses a problem for the subjectivist requires further examination. Parfit (see, e.g. part 7 in his

2017) presents arguments against what he terms analytical subjectivism, but these arguments do not aim to demonstrate that analytical subjectivism is inherently incoherent (see, also Parfit 2011a, 1:72–73). Thus, it seems that, at least *prima facie,* it is not 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 appears valid, considering that (M) or a similar concept, such as Williams' notion of a sound deliberative route, aims to clarify the concept of a reason, as understood by the phrase 'counting in favor of". The distinction between object-based and subject-based theories of reasons lies in how they explain the 'counting in favor of' relation, rather than the more substantial question of which facts (states of affairs, their features, etc.) precisely count in favor of what. The response to the latter question will depend on how we interpret the idea of a sound deliberative route (for subjectivists), or on more direct intuitions about the intrinsic value of things (for objectivists) (for discussion, see Smith 2009).

Parfit may concur with the above line of reasoning, since, when giving the incoherence 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" (see Parfit 2011a, 1:72–73).

The question now arises: if subjectivists acknowledge that the concept of a reason is primitive, meaning it cannot be defined in terms of, for instance, a rational deliberative route, does this render their theories incoherent? Can analytical subjectivists alone sidestep the incoherence argument? I do not believe this is the case. Allow me to elucidate why.

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, we can assert that the statement "There is a reason to do X" is extensionally equivalent to the statement "there is a rational deliberative route that could lead one to do X". Affirming the extensional truth-conditions for these two statements does not imply that the concept of a reason reduces to or shares the same meaning as the concept of a rational deliberative route.[25] Thus, a person who is competent regarding the concept of a reason does not have to *a priori* recognize, simply on the basis of their competency with the concept of a normative reason, that all that is captured with the concept of a normative 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). Additionally, in the context of a rational deliberative route framed in subjectivist terms, the only presupposition required is that, irrespective of the reasons involved, they are reasons due to some association with agents' rational capacities and not inherent values in certain states of affairs. The core idea of this perspective is that objectivists and subjectivists can both adopt the same concept of a normative reason. They can even assert that this concept is normatively irreducible. However, they may still differ on substantive matters regarding why certain reasons apply to certain agents or what qualifies them as reasons, among other aspects (for discussion, see Street 2017).

Consequently, Parfit's incoherence argument still falls short even when interpreting (M) as non-analytical. The failure lies in

25 For example, Christopher Peacocke gives the following influential 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, 2).

Parfit's presupposition that adopting the concept of a reason as indefinable entails a commitment to an object-based foundation for the "counting in favor of" relation. However, this assumption is not necessary, as we can share a common understanding of the notion of a normative reason (at least in a pre-theoretical sense) without necessitating shared 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 consider two individuals, Joe and Mary. Mary was raised by parents who adhered to a realist philosophy, considering colors as objective and intrinsic features of objects. In contrast, Joe's parents embraced a response-dependentist view, viewing colors as not intrinsic but as dispositions of objects that can induce color experiences in perceivers under certain conditions. Despite their divergent background theories, when Joe and Mary discuss colors, they understand each other perfectly well; from their perspectives, both are adept at using color concepts. So, when Joe asks Mary for the "red cup", Mary hands him the cup, and when Mary describes a house's color as "hideous", Joe concurs, as he shares her dislike for houses painted in vivid green and red.

It appears evident that Mary and Joe are proficient in applying color concepts, and most of the time, their discussions about colors revolve around the same referents. The only scenario in which they might not agree is when the nature of color is explicitly debated. Due to their distinct upbringings, Mary views colors as intrinsic features of objects, while Joe sees colors as response-dependent properties. Whether one believes that background ontological theories should partially shape a concept and our competence with it or not, it seems reasonable to assert that, at least pre-theoretically—before examining the ontology of colors—Mary and Joe share the same concept of color.

Returning to our discussion of the concept of a normative reason, it is worth noting that the irreducibility, indefinability,

or primitiveness of this concept does not necessitate commitment to any particular ontology of reasons. This parallels the way the indefinability and primitiveness of certain color concepts do not tie us to a specific color ontology (Street 2017). If we accept that the common beliefs about normative reasons presented in Table 1 are something theories of reasons should accommodate, then, at least on the surface, it appears that object-based and subject-based theories are on the same pre-theoretical ground.

This line of reasoning should help us understand that even if we interpret Parfit's claim (M) as substantive, we can still argue that what provides a basis for believing that our reasons align with (M) is (O'). The primary disagreement between objectivists and subjectivists regarding reasons lies in the ontology of reasons, rather than the question of which reasons exist.

2.6.3 Why idealize?

At this juncture, I would like to consider another line of thought that may be driving Parfit's intuitions underlying the incoherence argument. This will set the foundation for what will be discussed in the next chapter.

One justification for Parfit's assertion that only statement (O) can account for the validity of statements (M) and (N) is the perspective that if normative reasons are not mind-independent and derived from intrinsic features of things, the introduction of idealization conditions into the discussion about reasons would be rendered meaningless.

Indeed, this is the worry that is most compellingly articulated by David Enoch (2005). The worry is that if we cannot give some kind of non-*ad hoc* reason for introducing idealizations into subjectivist accounts, then the threat is that subjectivist theories (that endorse some version of idealization) would be dangerously unstable. The primary role of idealization is to enhance our epistemic standpoint, particularly in non-optimal epistemic conditions. Essentially, idealization aids in rectifying these epis-

temically challenging circumstances. For instance, if we struggle to discern the time on a clock from a distance, getting closer would place us in a better epistemic position. This is because an individual standing closer to the clock would be in a superior epistemic position to the same person viewing it from a distance. In such instances, this form of epistemic idealization is sensible because we are attempting to ascertain a fact that exists independently of the tracking process.

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 as captured by (O). However, it appears that 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 is that according to statement (M), reasons are provided by what a rational deliberator would desire, decide, or aim to do. However, in subject-based theories, where reasons hinge on the preferences of a rational deliberator, the idealization procedure intended to track normative reasons is not entirely independent of the tracking process. Consequently, some argue that subject-based theorists lack a compelling justification for incorporating the idealization procedure into their account of normative reasons (Enoch 2005, 764–65). This leaves us with a question: if someone adopts a subjectivist stance on reasons, what justifies the incorporation of idealization?

Considering my conviction that a naturalist account of the ontology of normative reasons naturally aligns with subject-based theories that incorporate some form of idealization about what we would desire if we were rational, I am now tasked with explaining why subject-based theories are entitled to employ idealization in our accounts of normative reasons. Addressing these questions will be the focus of the next two chapters.

2.7 Summary

In this chapter, in alignment with Parfit (2011a), I made a distinction between two models of normative reasons: object-based and subject-based theories. I contended that, from a naturalistic standpoint, subject-based accounts are a more fitting choice. Subsequently, I addressed notable objections to subject-based theories and explored potential responses to these critiques. The conclusion of the chapter ended with a brief examination of the "why idealize" objection raised against a plausible form of subject-based theory presented by Enoch (2005). This discussion sets the stage for the forthcoming exploration of normative reasons from a naturalistic perspective in the next two chapters.

3 Response-dependence and the problem of idealization

3.1 Introduction

In the preceding chapter, I mentioned the potential connection between Parfit's (2011a) *incoherence argument* and Enoch's (2005) critique of subjectivists relying on idealization conditions without a robust justification for these conditions in the absence of a belief in subject- or mind-independent normative facts. The aim of this chapter is to show that, in principle, a version of a subject-based theory that I will refer to as a response-dependence view of normative reasons can motivate idealization that is not in some objectionable way *ad hoc*.

In the chapter, I proceed as follows. In the next section, I provide a model of a subject-based theory in the form of a response-dependent theory of normative reasons. Then I present Enoch's why idealize challenge to this view of normative reasons.[26] In the following two sections, I explore two ways of responding to Enoch's challenge. One way involves a revisionary stance on the ontological commitments of the normative discourse about reasons. The second route involves the denial of Enoch's contention that our normative discourse is implicitly committed to a realist ontology. The overarching claim is that our normative discourse only presupposes a possibility of misrepresentation. However, this feature of normative discourse does not favor robustly objectivist accounts of normative reasons over response-dependent ones. Therefore, this feature can be freely used by a proponent of a response-dependence account of reasons to answer the question of why to idealize.

26 David Sobel (2009, ft. 3), in his discussion of Enoch's (2005) paper, lists other authors who raise similar objections to subjectivist theories of normative reasons.

3.2 A response-dependence account of reasons and Enoch's challenge

Enoch's challenge can be framed in terms of normative reasons. As we saw in chapter 1, it is common to think about normative reasons as facts that *count in favor of* something. For example, the fact that eating an ice cream would give me gustatory pleasure counts in favor of my eating the ice cream. Subjectivists and objectivists about normative reasons can be differentiated by their views on what grounds this *counts in favor of* relation. As we saw in the previous chapter, objectivists contend that the fact that something counts in favor of believing or doing something is a completely objective and mind- or response-independent truth. Subjectivists contend that truths about normative reasons are determined in relation to some facts about agents, such as their desires, preferences, motivations, values, beliefs, and so on.

Subjectivist theories have several appealing features. For instance, they can easily accommodate variability in reasons among different agents. The fact that there will be dancing at a party tonight is a reason for Ronny but not for Bradley to go to the party (M. A. Schroeder 2007). What intuitively explains such a difference in reasons is the fact that Ronny has a desire for dancing, while Bradley does not have a similar desire. Subjectivist theories can also easily account for the motivational relevance of reasons (Williams 1981). If facts provide normative reasons in virtue of affective and other motivational facts about agents, then it seems clear how those facts would motivate an agent.

Furthermore, such theories are often seen as more in line with a naturalistic worldview (Enoch 2005). If normative facts are not grounded in facts about agents, then there is a standing worry about how to explain their nature in a non-mysterious way. In addition, if facts about normative reasons are robustly mind-independent and non-causal, then we might wonder how we come to know these facts. And more importantly, how can we be sure that our deeply entrenched judgments about norma-

tive reasons are in fact true, given that we are biologically evolved creatures with limited cognitive powers (see Street 2006)? Subjectivist theories seem to have an upper hand in this respect, given that they ground normative reasons in (relational) facts about agents. In what follows, I develop a response-dependence model of normative reasons that accommodates the aforementioned considerations and explore its prospects for answering Enoch's challenge.

A typical account of response-dependent properties starts with a characterization of concepts that pick out those properties. The characterization standardly involves a biconditional of the following form (see, e.g. Johnston 1989):

> (GD) O is F iff O has a disposition to elicit a response R from such and such a person P under conditions C.

The order of determination is from right to left; O is F *in virtue of* O's having a disposition to elicit R under C. A response-dependent property, then, is that property which is picked out by a response-dependent concept.[27]

A response-dependence account of normative reasons could run as follows:

> (RD) The fact that p is a normative reason for X to F iff X is disposed to F on the basis of p, in conditions where X is rational.[28]

Here, F stands for a general action verb. It is intended to be maximally inclusive. Depending on the specific view, it could include a performance of an action and/or producing mental responses, such as desires, beliefs, valuing, inferring something on the basis of p, and so forth. In addition, a disposition to F does not necessarily involve responses that are consciously or deliberately made on the basis of p. It is intended to also encompass the idea that reasons can be determined by spontaneous and non-deliberative

27 For further qualifications of this claim, see López De Sa (2013a).
28 For other examples of this type of response-dependence accounts of reasons, see, for instance, Goldman (2009) and Williams (1981).

processes. For instance, the fact that X is hungry gives her a reason to eat something because, in normal conditions, she has a disposition to respond to this fact in this way. However, this disposition is presumably not based on X's deliberative processes.

The conditions of rationality can be spelled out differently, depending on one's conception of norms of rationality (see, e.g. Parfit 2011a, 1:61–63). Not to prejudge substantive views of rationality, we can think of it in the following broad terms. A rational person is someone who avoids actions, motivations or thoughts that would be self-defeating with respect to her background concerns and motivations (see Goldman 2009, 45–82). In addition, a rational person is someone who has an ability to reliably track information from her surroundings and to appropriately employ her reflective capacities.[29]

Following Ralph Wedgwood (2007b, 88), here I understand dispositions as functions from stimulus to response conditions. A standard example is the brittleness of a glass. If a glass is struck (the stimulus condition) it breaks (the response condition). However, an ascription of dispositions involves a reference to normal conditions. If a glass is struck and it does not break, it does not necessarily follow that it lacks the property of being brittle. Rather, this could indicate that the stimulus conditions are not normal (maybe it was not struck with enough force or an 'angel' made a protective belt around it). In (RD), the stimulus conditions refer to a fact that provides a reason. The response conditions refer to whatever is captured by "F-ing". The normal conditions are, among other possible things, provided by the conditions of rationality. If a person does not respond to a fact that p by F-ing, it does not necessarily follow that p is not a reason to F. It could indicate that the conditions are not normal, that is, that the person is not rational or the conditions inhibit the manifestation of

29 For discussion of the value and purpose of deliberation for rationality, see Arpaly and Schroeder (2012, 230–36).

rationality.

A response-dependence view of reasons is supposed to capture the idea that although reasons can be provided by response-independent facts, their normative status as reason-giving is constituted by some dispositions of rational agents. As such, this view accommodates the variability of reasons and their motivational impact by relativizing reasons to cognitive and conative dispositions of agents. This account is in line with a naturalistic worldview because it grounds normative facts in relations that agents bear to the world (for further discussion, see Sun 2022). Furthermore, it avoids skeptical worries, given that such facts can be discovered and tested by ordinary methods that we employ in discovering truths about ourselves and our surroundings.

Now I turn to Enoch's challenge. The question is: why should a response-dependence account of normative reasons include idealization, or in my model, conditions of rationality? Many objectivists and subjectivists agree, for instance, that not every desire provides or in some other way determines a normative reason for action (see, e.g. Parfit 2011a; Smith 2004; Williams 1981). According to Enoch, "a natural rationale" for introducing an idealization condition into an account of normative reasons, for instance, "would be to claim that the relevantly ideal conditions are the conditions needed for a reliable tracking of the relevant facts" (Enoch 2005, 761–72). However, in normal circumstances, the relevant facts are response-independent. Since, *per* (RD), normative facts are in some sense dependent on the agent's responses, we cannot rely on the natural rationale to justify idealization. Enoch illustrates the application of "the natural answer" with the following example:

> 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, 762)

According to Enoch, in this case it makes sense to use idealization because our capacities and resources in the current epistemic situation might not reliably indicate what really is the case. Idealization refers to better epistemic conditions, through which one acquires a belief about what really is the case.

The alleged problem for response-dependence views is that since they do not posit response-independent normative facts, this epistemic reading of idealization is unavailable to them. In fact, the idealizing conditions in (RD) aim to capture the metaphysical grounding of the normative and not the epistemic procedure for tracking facts, regardless of whether they are response-dependent or independent. Thus, "the natural rationale", which relies on epistemic considerations, is not natural at all in the context of providing a metaphysical account of normative properties. So, what would constitute a more appropriate vindication of idealization in the present context?

I maintain that a plausible answer is that our normative discourse presupposes the possibility of being mistaken in our normative judgments (for a similar view, see Prinz 2006, 35). To use Williams' (1981) notorious example, Jones wants to drink gin and tonic and thus orders it, but the waiter, unbeknownst to Jones gives him a glass full of petrol, which under no circumstances does Jones want to drink. Since Jones does not know that the glass contains petrol, he drinks the stuff. Intuitively, Jones does not have a reason to drink from the glass. He might have thought that he had a reason, and that would explain why he drank the petrol. But, in fact, had he acquired the information about the petrol in his glass, he would have realized that his judgment about the reasons he had was false. At least in the case of normative reasons, this possibility of misrepresentation is borne out

by introducing some requirement of rationality and possession of correct information that an agent's responses need to satisfy.[30]

Enoch seems to be aware of this type of defense:

> Somewhat more generally, we think that we are fallible in our normative judgments and that there is room for genuine normative advice and for coming to see that we were mistaken about our reasons. (Enoch 2005, 769)

However, he thinks of this defense as relying on our practice of justifying and criticizing statements about reasons. According to this line of argument, "[t]he hope is, then, that from our practices of justifying normative claims a rationale for idealization can be extracted" (Enoch 2005, 769).[31]

30 A word of caution is necessary here. This explanation of why to idealize is different from the explanation that Enoch terms the "extensional adequacy" (Enoch 2005, 766–69). If I understand Enoch correctly, the explanation based on extensional adequacy refers to the idea that idealization can be justified by a response-dependence theorist's "desire" to accommodate intuitive judgments about what reasons we have, and thus avoid possible intuitive counterexamples to her theory. Idealization, on this view, is to secure extensional adequacy of the normative implications of a response-dependence account and our intuitive judgments about what there is a reason to do or believe. If this were the only justification for introducing idealization into a response-dependence account, then I agree with Enoch that it would be *ad hoc*. However, according to my explanation, (RD) introduces idealization to capture the possibility of misrepresentation and not to accommodate specific judgments about reasons we might have. Later in the paper, I will argue that this feature of our normative discourse about reasons is neutral with respect to the ultimate ontological status and specific contents of normative truths.

31 This is not exactly right. The explanation that I am alluding to, that normative discourse presupposes the possibility of being in error, is not based on considerations relating to how we justify normative claims. My explanation is based on semantical and/or ontological considerations. Concepts of response-dependent properties, like other concepts, are such that you can misapply them. In addition, standard models of response-dependent properties rely on the concept of a disposition. The nature of dispositions, however, involves a reference to normal (ideal, counterfactual, *ceteris paribus*, etc.) conditions, which guarantee their manifestation (Wedgwood 2007b, 88). Thus, strictly speaking, (RD) and its reliance on some kind of idealization, should be probed from the perspective of its aim, which is, in the first instance, to provide an ontology of reasons and not their epistemology. The epistemology can be extracted on the basis of its ontology, but this is not the main focus when we try to provide truth conditions for claims about reasons. This point will become important later in the paper.

According to Enoch, this explanation is again unavailable to an (RD) supporter.

> What best explains our justificatory practice is rather our (perhaps implicit) belief, false though it may be, that, say, conditions of full imaginative acquaintance are conducive to the reliable tracking of an independent order of value-facts. And the idealizer cannot consistently require an explanation why it is a good justificatory method, because she believes it is not: she believes that there is no independent order of value-facts that our epistemic methods reliably track. (Enoch 2005, 775)

The main claim is that an (RD) idealizer cannot rely on our justificatory practices to vindicate idealization because idealization as a justificatory method is best vindicated against a background of a response-independent normative ontology.

At this point, it is useful to introduce a distinction between revisionary and non-revisionary response-dependence views. Hallvard Lillhammer distinguishes between the two views as follows:

> The analytical dispositionalist claims that the response dependence of normative reasons is a conceptual truth which can be read off from constitutive commitments implicit in commonsense ethical discourse. (...) The revisionary dispositionalist, by contrast, claims that normative reasons should be construed as response dependent regardless of the conceptual commitments embodied in common sense ethical discourse. (Lillehammer 2000, 174)

The usual motivations driving revisionary dispositionalism are the metaphysical and epistemological worries accompanying the robust realist construal of our normative discourse. For instance, Mackie (1977) argued that robust realism about morality is committed to postulating metaphysically *queer* entities.

The important thing for the present discussion is that Enoch (2005, sec. V) maintains that his argument against the plausibility of response-dependence views of normativity does not apply to revisionary accounts. Enoch (2005, 770–74) argues that our normative discourse is implicitly committed to robust objectivism. However, he maintains that if one can show that there is

something wrong with the default, response-independent view of normative discourse, then response-dependence idealizers are "off the hook". Nevertheless, and more importantly, he contends that "the revisionist idealizer cannot rely on the natural answer any more than the nonrevisionist idealizer can" (Enoch 2005, 786). The reason, again, seems to be based on the claim that the purpose of idealization is to track "an independent order of value-facts".

In the next section, I grant Enoch's premiss that normative discourse is committed to robust objectivism. Against this background, I explore the prospects of a revisionary response-dependence account providing a plausible answer to the question of why to idealize. I argue, that in principle, a response-dependence theorist can provide the "natural answer" to "Why idealize?". I will argue for this conclusion by relying on the case study of color perception.

3.3 Response-dependence about color and the natural answer

It seems that color, phenomenologically speaking, is presented to a normal observer as an intrinsic property of material objects (see Clark 2000, 13–14; Giere 2006, 25). However, empirical research on color perception provides good reasons for thinking that colors are not intrinsic, objective properties of external objects (see Palmer 1999, 95).

There are at least two related reasons for thinking this.[32] The first is the structure of the hue dimension of the color space. The human visual system differentiates four color hues: red, green,

32 In framing the present section on color perception, I rely on Ronald Giere (2006, ch. 2). However, a caveat is in order. There are some philosophers who argue that data on color perception could be interpreted in a way that is compatible with realism and objectivism about colors (see, e.g. Byrne and Hilbert 2003; Tye 2002). Despite the scientific evidence, I do not want to commit myself to any strong conclusions about the ontology of colors. For the purposes of the analogy, it is enough that there is a respectable view according to which colors are response-dependent properties.

blue and yellow. What is important about the hue of the color space is that structurally it has a circular form.[33] If colors could be reduced to response-independent physical properties, then the natural reductive base would be physical properties of the spectral reflectance of a surface and wavelengths of the light that gets reflected from those surfaces (Giere 2006, 26; see, also Hardin 2003).

If that were the case, then we would be able to say, for example, that the object O is red because it reflects light of the wavelength X. However, such identifications of "perceived hues with single wavelengths" are standardly not possible because the structure of wavelength is linear as opposed to the circular structure of the hue dimension of the perceived color (Giere 2006, 18).

The second reason is closely connected to the first and it involves the phenomenon called *metamerism*. Metamerism is a phenomenon in which the light of *different* wavelengths can produce the *same* phenomenal color experience in a normal observer.[34] Moreover, across individuals, different color experiences can be produced by the same combination of wavelengths presented in the stimuli (Clark 2000, 11). 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, nonetheless, remains the same for all normal observers.

For every normal observer, thus, 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

33 In geometric terms, red-green hues form one continuous axis and blue-yellow form the other. Together, they form a hue circle (see Giere 2006, 17–18; see, also Palmer 1999, 98–99).

34 All combinations of wavelengths that have the same impact on the visual system are called *metamers* (Clark 2000, 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 (Clark 2000, 10).

will be determined by its place in the color space.[35] Because of the circular form of the hue of the colors, for every normal observer, green will be characterized negatively in relation to yellow and blue. That is, it can be characterized as not being yellowish and not being bluish. Other colors, such as orange, for example, would be characterized in positive relation to other colors, that is, for every normal observer, orange will be identified as a color that is between red and yellow. What is important to mention here is that even though different stimuli can produce different color experiences in different subjects, once we identify which combination of wavelengths produces 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 be, in the structurally relative sense, the same across normal individuals that are members of the same species (see Clark 2000, 11–12).

Given this scientific evidence, a response-dependence view might provide a necessary revision of our commonsensical color ontology that preserves its important aspects. A response-dependence view of colors could be spelled out as follows:

> (CD) O has color C *iff* O tends to elicit phenomenal experience E from persons with normal visual system P under viewing conditions C.

Here, again, the biconditional should be read from right to left; O has color C *in virtue* of its tendency to elicit E under C (see, e.g. Johnston 1992; López De Sa 2013a; Miščević 2004).

The seeming objectivity of colors is mostly related to the phenomenon of color constancy.[36] It enables us to "distinguish

35 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 an experience of a greenish red or bluish yellow color (see Palmer 1999, ch. 3).

36 Color constancy refers to "the stability of perceived object color across changes in viewing conditions" (W. Wright 2013, 435). We perceive, for example, an apple as red, whether the apple is placed in deep shade, or is illuminated by white sunlight.

between color *appearance* and color *reality*. (...) [W]e think we know someone's red BMW is really red even though it does not appear red at night in a parking lot illuminated by sodium vapor lamps" (Giere 2006, 24). To the extent that normal color-perceivers think of colors as intrinsic properties of objects, a proponent of (CD) would saddle them with erroneous beliefs. However, as emphasized by Nenad Miščević (2007), one of the positive sides of adopting (CD) is its ability to charitably interpret naive cognizers. In fact, if we grant the reasons for revision, (CD) is charitable to naive cognizers because, by reinterpreting their color discourse as referring to response-dependent properties, it saves the rationality of naive cognizers and maximizes "truth-likeness of [their] views" (Miščević 2007, 216). In addition, (CD) sustains the difference between appearance and reality by including normal conditions that determine when color concepts are correctly applied.[37]

Now we can ask what the reason is for idealization in the case of color. Well, it seems that, depending on how we construe the question, there are least two distinct ways of answering it. One answer is related to general reasons why we would accept (CD) in the first place. In this section, I offered reasons that include scientific evidence and our tendency to minimally revise our commonsense theories in a way that is charitable to the beliefs of naive cognizers. If we grant these reasons for revision, then idealization, namely dispositions and their conditions of manifestation, comes with the territory.

The second way of construing the question is more akin to

37 We can even reasonably speculate about what fixes the normal conditions. Giere (2006, 31) writes that color vision has evolutionary selective advantages. Color enables organisms to identify objects as their conspecifics, potential mates, edible, and so forth (see Mollon 1989). Given the human evolutionary story, the recognition and identification of objects by their color is most reliable during daylight. Thus, a reasonable supposition is that a necessary condition for determining normal conditions in (CD) involves the experience of phenomenal color, caused by objects that are illuminated during daylight.

Enoch's epistemological challenge. Why idealize if you are not an objectivist about colors? Once we grant the plausibility of (CD), though, the epistemic justification seems to be straightforward. In the case of colors, we idealize to capture the normal conditions, where our responses authorize the application of a certain color concept. For instance, a person looking at a BMW illuminated by orange light might be inclined to judge that the car is blue. However, realizing that she is not in normal lightning conditions, she abstains from that judgment. The idealization in her case enables her to reach the conditions where, according to (CD), her visual responses determine the correct judgments about the car's color. In this case, the motivation for idealization seems to be perfectly in line with the underlying reasons for adopting a response-dependence account of colors.[38]

Similar considerations apply to our response-dependence account of normative reasons.[39] There are various reasons for adopting a response-dependence view of normative reasons. Analogously to the case of colors, contemporary naturalists in metaethics think that there are considerations that make robust normative objectivism incompatible with a scientifically grounded picture of the world. Some argue, for instance, that sociological studies of diverse cultures and cognitive science of morality indicate that moral properties are best construed as being grounded in our concerns and affective responses (e.g. Prinz

38 Thanks to Michael Smith and Luca Malatesti, respectively, for indicating the need and helping me to phrase the main point of this paragraph more clearly.

39 The analogy with colors has its limits though. For instance, we can expect that color phenomenology will not display great variance among individuals, since it depends on perceptual mechanisms that are mostly uniform among humans. However, we can expect greater variance in the content of reasons, since they depend on higher order mental processes (preferences, desires, reasoning styles, etc.) that seem to display more variance among individuals. These differences can be set aside since the argument relies only on the structural similarities between the two accounts. For a discussion of similarities and differences between response-dependence views of color and value, see López De Sa (2013b).

2006). Others rely on evolutionary considerations that seem to put pressure on the idea that normativity is something objective about which we could have knowledge (Joyce 2006). Street (2006) has argued on evolutionary grounds that if we have some normative knowledge, then that knowledge should be construed as being of mind-dependent facts. Whatever the merits of these arguments, they at least provide prima facie reasons for thinking that the ontology of normative reasons cannot be robustly objective.[40] Analogously to the case of color ontology, it could be argued that (RD) provides a plausible minimal revision of our ontology of reasons that is charitable to naive cognizers. There is nothing in this explanation that would make the account and its idealizing components (i.e. dispositions and their manifestations) *ad hoc*.

Furthermore, we have, again, two ways of responding to Enoch's challenge. The first response depends on the original motivations for adopting (RD). If it is granted that (RD) provides a plausible revision of our ordinary reasons-ontology, then the explanation of idealization will be grounded on the nature of dispositions and their normal conditions. The second response will invoke the idea that by idealizing we aspire to put ourselves in the conditions, which according to our theory, enable us to track our reasons.[41]

Nevertheless, Enoch (2005, 786) contends that a revisionary response-dependence account of reasons cannot provide "the natural answer" to the question, "Why idealize?" In fact, Enoch claims that his challenge generalizes to all response-dependence accounts regardless of their subject matter:

> the challenge—that of coming up with a rationale for the idealization that is consistent with the philosophical concerns underlying the relevant response-dependence view and that is not objectionably ad hoc—applies across the board. [...] Furthermore, the unavailability of the first, natural answer to the why-idealize

40 In fact, I will defend a version of such an argument in Chapter 5.
41 Enoch (2005, 770) seems to condone this point.

> challenge is also independent of subject matter. So long as the relevant idealizer thinks of her view in the way characterized in Section II above—where the relevant truth depends on our relevant responses, and not the other way around—she cannot motivate the idealization by considerations regarding the accurate tracking of an independent truth. (Enoch 2005, 781–82)

However, it is not clear why a proponent of (CD) cannot use "the natural answer" to "why idealize" when the question is understood as epistemological. In this quote, Enoch relies on the claim that "the natural answer" involves a commitment to the tracking of independent truths. It is not clear what the argument for this claim is. At the beginning of the paper, Enoch claimed that the natural answer can be provided just in "cases where the relevant procedure or response is thought of as tracking a truth independent of it" (2005, 764). For instance, he claims that "when we think of one thing (the watch reading) as a reliable indicator of another (the time), we think of the latter as independent of the former" (2005, 763). Even granting that we normally think of idealization as an epistemic procedure that involves tracking truths that are independent of it *as a procedure*, it does not follow that the *tracked truth itself must be response-independent*. This would involve an invalid inference from epistemological premisses to ontological conclusions.

Let me illustrate this point with an example. Let us say that for a naive cognizer A, red is whatever produces in her an experience of red in normal conditions. For A, then, a good epistemic procedure for determining whether an object is red is to observe that object under normal viewing conditions and see whether her experience of that object will involve a sensation of red. Let us suppose that after A learns about the scientific evidence about color perception, she adopts (CD). Would A's confidence in her original procedure for determining the color of an object decrease after this theory change? I am inclined to think that it would not. After all, her new theory of color would vindicate the same epistemic procedure. This, in turn, implies that her epis-

temic procedure (i.e. idealization) was not committed to tracking response-independent truths to begin with, even if it seemed so to A before the theory change.[42] Note, furthermore, that A could, even after the theory change, think about idealization as tracking facts that are independent from the procedure. Under the epistemic mode of presentation, idealization would be thought of as having a tracking function of facts, regardless of their ontology. Under the ontological mode of presentation, idealization would be thought of as a constitutive element of some properties or facts. In this sense, A could think of the natural answer for idealization as tracking something that is (conceptually, but not necessarily ontologically) independent from it.

If this point works for response-dependence about colors then it should, *mutatis mutandis*, apply to response-dependence about normative reasons. Thus, the purported conclusion of this section is that a revisionary response-dependence account can provide "the natural answer" to "why idealize". In the next section, I explore the prospects of a non-revisionary response-dependence account of reasons for supplying a plausible answer as to why to idealize.

3.4 Prospects for a non-revisionary response-dependence account of reasons

Enoch grounds his "why idealize" objection to non-revisionary idealizers in the claim that ordinary normative discourse and its justificatory practices are committed to robust realism because a commitment to this ontology best explains our practice. Here is an indicative quote:

> Regardless of how good her (metaphysical, epistemological, or whatever) reasons are for denying a more robustly realist view of the relevant normative truths, still these are not reasons to deny

42 In footnote 30, Enoch (2005, 773) considers a similar reply. However, he fails to consider it as a potential justification that a revisionary response-dependence theorizer might use to supply "the natural answer."

> that our justificatory practices are committed to some such realism. (...) What best explains our justificatory practice is rather our (perhaps implicit) belief, false though it may be, that, say, [ideal] conditions (...) are conducive to the reliable tracking of an independent order of value-facts. And the idealizer cannot consistently require an explanation why it is a good justificatory method, because she believes it is not: she believes that there is no independent order of value-facts that our epistemic methods reliably track. (Enoch 2005, 773–74)

So far, I have been following Enoch in treating "the natural answer" and the explanation based on justificatory practices as conceptually different possible answers to "why idealize". However, they really come down to the same thing. If we ask what justifies the claim that "the natural answer" involves tracking of response-independent facts, the only salient answer seems to be that this is what best explains our practice of using idealization. Thus, in what follows, I maintain that if a response-dependence account can rely on our justificatory practice to vindicate idealization, then that account comes as close as it can to providing "the natural answer" to the "why idealize" challenge.

To bolster his claim that our justificatory practices track response-independent facts, Enoch relies on a thought experiment. Suppose that there is a view of religious obligation called the Ideal Prophet Theory (Enoch 2005, 770–71). Proponents of such a theory, given their naturalistic inclinations, deny the existence of God (including other supernatural entities). They propose the following formulation of their theory:

> (IPT) Action A is religiously required iff it is sensed-as-required by a prophet in ideal conditions. (Enoch 2005, 771)

Proponents of (IPT) specify, in a naturalistically respectable way, what sensing-as-required means and what the ideal conditions are.

How could they respond to Enoch's challenge? The most sensible explanation would rely on the justificatory practice embedded in this religion. They could claim that what vindicates

idealization as a good method of coming to know what is religiously required is the fact that this procedure is partly constitutive of the property of *being religiously required*.

Nevertheless, Enoch claims that there is a better explanation for idealization in the vicinity that does not favor (IPT).

> The problem I want to focus on stems from the gap between the commitments of the Ideal Prophet Theorist and those of the participants in the relevant religious practice. (...) This gap makes the Ideal Prophet Theorist's argument into a non sequitur: the best explanation of the relevant justificatory practice is not in terms of the theorist's metaphysical beliefs (...), but in terms of the participants' ones. Participants believe in God (or in other supernatural facts), and so the best explanation of the relevant part of religious practice is the obvious one: participants believe that the relevantly ideal conditions are the conditions best suited for the tracking of an independent order of facts regarding religious obligations (...). (Enoch 2005, 772)

According to Enoch, the normative idealizer's appeal to "our justificatory practices fails in an exactly analogous way" (2005, 773), implying that the beliefs of the ordinary participants in the practice of justifying and refuting reasons are committed to a robustly realist ontology.

It certainly makes sense to think that what justifies religious practice is the participants' belief in God. Thus, it is not clear why or how anyone could reconstruct a religious practice in a way that would eliminate its constitutive element (i.e. God). It seems to me that at most, the (IPT) could offer a new grounding for morality which is derived from, but not constituted by the religious practice. However, for normative discourse about reasons, it yet needs to be shown that response-dependence views eliminate the constitutive element of that practice.[43]

In fact, there is evidence that metanormative beliefs of ordinary people are not clearly committed to robust realism; rather, they exhibit a pluralistic pattern (for discussion, see Pölzler

43 Enoch (2005, 774) seems to condone this point as well.

2018; 2023). For instance, Hagop Sarkissian and colleagues (2011) found that people's beliefs about the objectivity of moral facts are culturally bound and correlate with how open people are to alternative perspectives.[44] Geoffrey Goodwin and John Darley (2008) and Jennifer Wright, Piper Grandjean, and Cullen McWhite (2013) found that people's patterns of metaethical intuitions vary inter- and intrapersonally. For instance, classifications of issues as moral, social, personal, etc. vary between different people. To give a specific example, around 51% of people think that the issue whether first trimester abortion is right or wrong is a moral issue, while 41% think it belongs to the domain of personal choice (J. C. Wright, Grandjean, and McWhite 2013, 340). In addition, even when a person classifies an issue as moral, his or her intuitions about the objectivity of judgments regarding the issue vary as well. Concretely, around 60% of participants classify stem cell research, assisted suicide, and first trimester abortion as moral issues. Nevertheless, most of them think that if people disagree about these issues then this is a matter of personal opinion, indicating a non-objectivist stance on the issue. Thus, the overall opinion is that

> [participants] viewed the rightness/wrongness of some moral actions as being determined by the beliefs, norms, and values of the individual acting—or, less frequently, the community in which the individual acted—and the rightness/wrongness of other moral actions as being grounded in more objective bedrock, citing as the source the harm caused, matters of justice, the sanctity of life, self-evident truth, and so on. (J. C. Wright, Grandjean, and McWhite 2013, 353)

Wright, Grandjean, and McWhite (2013, 352–53) conclude that people's metaethical beliefs are neither strictly objectivist nor non-objectivist; rather, people exhibit a pluralistic pattern of in-

44 By this it is meant, for instance, that when exposed to information about different cultures, people exhibit non-objectivist intuitions. Relativism or non-objectivism refers to the idea that contradictory moral claims, in different contexts, can both be true (Sarkissian et al. 2011, 482).

tuitions. Given this variability in people's intuitions, it seems to be question begging to rely on the idea that what best explains justificatory practices is the robustly realist grounding of the normative domain (for discussion, see Pölzler 2023).

Enoch seems to anticipate this response but still rejects it as a plausible objection. He seems to think that explicit metanormative beliefs of the participants in the practice are not important for settling this issue. Rather, what is important are the implicit standards embedded in the practice itself. Here is the quote:

> what is relevant is not the explicit metanormative beliefs—much less the explicit metanormative statements—of participants in normative discourse. What is relevant, rather, are the deep metanormative commitments embedded (perhaps implicitly) in normative discourse and practice themselves. The fact that many sophomores (and not only them) express some subjectivist or relativist metanormative intuitions thus has very little weight in assessing the commitments of normative discourse. Indeed the attempt to motivate idealization by referring to our practice is, it seems, an attempt to motivate the idealization by reference to the standards implicit in our normative practice, not to whatever explicit metanormative beliefs participants may or may not have. (Enoch 2005, 773–74, ft. 31)

I find the contention that explicit metanormative beliefs of the participants are not important for determining the theoretical commitments of the practice rather puzzling. Especially because Enoch (2005, 773) himself earlier claimed that participants' *beliefs* are what matter for determining the metaphysical grounding of the practice. What if the participants in the Ideal Prophet Theory all claimed that for them the practice would make sense even if there were no God? Would that not show that their practice and the idealization it employs do not really depend on the existence of God? I would say that this consideration would provide at least some reason to think twice about the realist commitments of the practice. It seems to me that a proper dialectical move for a realist would be to try to explain why these metanormative beliefs are not important in this context. Thus, until a reason is provided

for why these beliefs should be disregarded, it seems to me that it would beg the question on the part of a robust normative realist to maintain that her ontology is what best explains the idealization in the domain of reasons.

It could be claimed that Enoch's analogy with (IPT) is not wholesale. Maybe in the normative domain, Enoch relies on the inference to the best explanation of the implicit standards of the practice, regardless of the participants' explicit metanormative beliefs. After all, their beliefs might not be the beliefs that they would hold after a suitable process of reflection. However, it is not clear what the grounds for the last contention are if they do not involve participants' metanormative beliefs. The only salient alternative is that, in general, idealization is best explained as tracking response-independent facts and hence presupposing a robustly realist ontology.

However, this claim is most certainly false. Consider an example. It seems to be a commonsensical view that the property of being funny is a flexible response-dependent property. It is flexible in the sense that what they will find funny can vary greatly among people (López De Sa 2013b). However, even such an overtly response-dependent property involves normal conditions for the correct application of the concept and, hence, idealization seems to be a good method for determining when things are funny for someone. Crispin Wright gives a taste of how this might work in practice:

> Basically, and obviously, the assertion condition for a comic statement is to experience amusement. But the warrants thereby conferred are open to defeat in a variety of ways (...). Avalanches, crying babies, drying paint, (...), a man pruning apple trees—none of these things could intelligibly be found funny without some very special stage setting. (...) The stilted and exaggerated stage mannerisms of a prima donna are not funny if you appreciate the magnificence of the music. (...) Then again, comic reactions can, of course, be merely badly informed and the claims they warrant correspondingly open to defeat by better information of no particular moral or aesthetic relevance. The politician's quip is

> not funny if you heard the heckler's question correctly, since the joke depended on an ambiguity which wasn't actually there. (C. Wright 1992, 100–101)

If generally the best explanation of the appropriateness of idealization is that the discourse is about a robustly realist domain, then the inference to the best explanation of why to idealize in the present case would indicate that funniness is a robustly objective property. However, being funny is normally construed as a response-dependent property. Thus, a further explanation should be provided of why in some cases idealization presupposes the existence of response-independent properties and facts, while in others it does not.

I propose an alternative unified explanation. What is similar to discourses that refer to properties such as being funny, being a reason for something, being red, and so forth, and more objective properties, is that they presuppose the distinction between appearance and reality. Thus, what explains the appropriateness of idealization is the possibility of being in error. The domain that presupposes the difference between appearance and reality is the domain to which idealization, in some form, can be appropriately applied. Since we can be wrong about whether some fact p is a reason to F even though we are disposed to F, and whether x is funny, even though x amuses us, these domains legislate the application of idealization. Thus, the supporter of a non-revisionary (RD) can provide "the natural answer": we idealize because we want to avoid error.

3.5 Conclusion

In this chapter, I have tried to disarm Enoch's "why idealize" challenge to a response-dependent account of normative properties. I have done this by exploring two lines of argument. According to the first, even if robust objectivism is the default view, possibly we have reasons to revise our normative ontology. The main point is that the function of idealization is to place us in conditions that

are metaphysically constitutive of the target properties (normal viewing conditions and rationality for colors and reasons, respectively).

The second line of argument involves questioning Enoch's supposition that idealization presupposes a robustly realist ontology. Experimental evidence indicates that a commonsensical normative discourse is ontologically pluralistic. Hence, it begs the question to rely on realist intuitions when discussing plausible explanations of idealization. Alternatively, I propose that what uniformly explains idealization in ordinary normative discourse is the commitment to the distinction between appearance and reality and the relative possibilities of error.

4 Idealization, deeper concerns, competing desires and non-parametric decisions

4.1 Introduction

In the previous chapter the main issue was to answer the question "why idealize" if we think that reasons have a subjective or response-dependent aspect. An important part of the answer was that we can be wrong about our reasons even if they are response-dependent. In particular, this seems to be clear when we think about the objective part of the counting in favor of relation; given our ends, preferences, concerns, and so on we want to know which facts are conducive to satisfying them. Nobody seems to be denying this much. However, human beings are creatures that not only discuss and evaluate the means to their ends but also the goals and ends that determine the reason-giving power of the relevant means. Thus, naturalistically oriented proponents of subject-based theories of reasons would also like to preserve the idea that our values, ends, and goals can be put under normative scrutiny. In this regard, idealization can be also used to provide an answer how ends might be evaluated from a naturalistic point of view. Furthermore, another reason for idealization involves capturing patterns of reasons that are grounded in the fact that people belong to a social species and need to get along to prosper. How this can be achieved and vindicated from a naturalistic point of view will be discussed in this chapter.

In this chapter, the discussion unfolds in the following manner. Initially, I explore how the conflicts in our numerous concerns and desires call for an evaluation not just of our instrumental desires but also the ones that underlie them, and what might be the role of idealization in this process. After that, I will explore how reasons can arise from interpersonal interactions and conflicts, the authority of which becomes evident from an idealized standpoint. Throughout this exploration, I will draw

on game-theoretical models to elucidate how idealization in such scenarios aids in tracking subject-based reasons emerging in interactions among human agents.

4.2 Human beings and deeper concerns

To provide a naturalistic vindication of the common view that not only means can be rationally evaluated, but also our goals that set the standards for instrumental reasoning, it needs to be recognized that we are creatures that have preferences, desires, ends, concerns, and so on that are often incompatible and compete for our cognitive resources. Put simply, a desire to do Φ essentially prompts you to do it. Nonetheless, it is important to note that not every desire necessarily mirrors your underlying or more profound concerns (Goldman 2009), the concerns that make up your identity, for example (Frankfurt 1988a; see also Jurjako 2017). Hence, when contemplating whether to fulfill a desire, we might be assessing its worth in light of the framework of our concerns that define our identity. Furthermore, our concerns are subject to potential misconceptions or necessitate exploration for a better understanding.

Moreover, when faced with incompatible desires, such as a desire to Φ conflicting with a desire to Ψ, it introduces a practical conflict that necessitates resolution for successful action. As indicated by Harry Frankfurt (1988a), conflicting desires, preferences, or ends create a context that requires us to adopt a stance regarding which desire, preference, or end we will embrace or identify with. This situation is inherently seen as one involving conflicts of valuations, and its resolution requires the provision of reasons (Korsgaard 2011). If this is the typical scenario encountered by us as human beings, it also appears to be ingrained in our nature to address these conflicts by utilizing our capacities for reflection and deliberation when considering available options. Effectively resolving these conflicts often involves establishing a minimally transitive order among desires, preferenc-

es, or ends and adopting a defined set of goals. These goals then serve as consistent constraints, offering reasons for actions within the context in which we have conflicting desires, preferences or commitments of another sort (Bratman 2007).[45]

Any endeavor to establish a consistent set of desires typically involves hypothetical thinking (i.e. idealizing), as our natural instincts and inclinations often generate conflicts between these desires. For instance, conflicts arise between the desire to maintain good health and the desire to smoke, between the inclination to procrastinate and the inclination to work, and between the preference for immediate rewards over larger future benefits, among others (see Ainslie 2001). These conflicts create challenges for us to effectively navigate the world in a manner that aligns with the individuals we are or aspire to be. The purpose of idealization in this context is to offer a vantage point from which solutions in harmony with one's profound concerns can be explored and implemented.

The concept of a defective desire holds a place within this framework. Those would be the desires that cannot be justified from the perspective of a person's deep or intrinsic concerns (Goldman 2009). The reason for prioritizing our deep concerns lies in their role in constituting our identities and determining the significance of certain things to us (see Frankfurt 1988b). Thus, we may assert that, in this regard, our deepest concerns and the desires arising from them possess a default evaluative authority. Naturally, their default status suggests the possibility of revision under pressure from factors like experience or deliberation. Nonetheless, their significance stems from the fact that they shape the agent's stance (see Frankfurt 1988a) or the perspective grounding the deliberation. In this sense, if a desire, such as aim-

45 See Frankfurt's (1988a, 170–71) 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.

lessly turning on radios or eating saucers of mud, fails to align with a person's deep concerns, she would lack a reason to engage in those activities because they would not hold importance for her. Upon reflection, such desires would likely be subject to revision or rejection (see, also Lenman 2005).

Desires can exhibit defects in other notable ways, particularly if they are self-defeating or contingent on factors that result in self-defeating actions, including errors in rationality (which encompass errors in our reasoning processes). For instance, a desire that conflicts with its own fulfillment, such as desiring both Φ and not ψ, when we know that ψ-ing is necessary to accomplish Φ, is likely to be perceived as defective or at least as non-optimal. In such a case, individuals would typically seek to eliminate, revise or at least abstain from acting on such desires as they cannot be satisfied. Additionally, if a desire, like the desire to eat mud, is contingent on the belief that regularly engaging in this behavior will maintain one's children's health, the desire may be considered defective due to its reliance on a faulty premise.

However, while the subject-based or response-dependent view of reasons allows us to understand the need for idealization and the concept of defective desires, it doesn't provide us with a systematic way to distinguish desires that are defective or irrational from those that are not. In this regard, unlike some authors, I do not think there are a priori grounds for asserting that individuals with peculiar desires, such as desiring to dedicate their lives to counting blades of grass, are irrational or possesses defective desires. On the contrary, considering how John Rawls (1971) originally frames such a scenario, such a person appears highly rational, adept at securing the means to execute her life plan. In this regard, not having strong a priori constraints about what kind of desires ground reasons leaves it open that different forms of lives, including conceptions propounded by the members of neurodivergence movements, can ground different conceptions of rationality that shape what people have reasons to do

and want. I think this is a positive aspect of an account of reasons (see, e.g. Lekić Barunčić 2021; Chapman 2023).

Nonetheless, the current perspective on reasons permits extensions that could potentially impose additional constraints on what might generally be considered a reasonable desire or the structure of concerns that would identify those reasons. Given that human beings are, to a large extent, social beings who need to engage in various interactions, cooperation, competition, and more, these social dynamics can impose constraints on individuals' desires, beliefs, and structures of concerns. This line of reasoning introduces a further noteworthy dimension in which idealization plays an important role in delineating the normative reasons a person possesses.

4.2.1 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 circumstances surrounding a decision 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 decides; 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 exam-

ple, when playing rock, paper, scissors[46] Smith's winning strategy depends on predicting the choices that Jones will make. However, Jones' winning strategy depends on predicting what Smith will play. Thus, 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 (see Verbeek 2008).[47] Suppose that my spouse 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 each of us chooses 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 another. Our reasons would depend on the mutual anticipation of our choices and, in effect, our reasons would refer to each other. Bruno 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

46 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.

47 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, 14). Thus, 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.

> 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, 74)

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 known as the hawk–dove game. This game was originally employed by the evolutionary biologist John Maynard-Smith for modeling interactions between organisms and their strategies that led to the evolution of cooperation (see Binmore 2007, 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, 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: 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 (adapted from Binmore 2007, 137).

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.[48] 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 (see Verbeek 2007, 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

48 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.

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.[49]

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 is the problem solved in real life? The theories 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, ch. 1; Verbeek 2007, 147–48).

For example, one stable pattern of interaction that seems to solve the problem includes the following strategy: if a person finds a property (e.g. land, forms of energy, commodity, etc.), 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.[50] For this strategy to become stable,

49 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, then 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.

50 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 indi-

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 established patterns 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.[51] 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.

viduals in the population were to conform to it.
51 Alternatively, maybe we should say that they are *expectation-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 deliberative point of view[52] 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 John Searle's (2001) objection to subject-based (or internalist) conception of reasons. Searle provides the objection in the form of an example.

> 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, 27)

52 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).

Thus, 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 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 **Figure 1** and **Figure 2**. The practice that gives rise to the expectations that are captured in Searle's example is a product of the evolution of human cultural 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 expec-

tations 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 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).[53]

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

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

53 Although, see Baccarini and Malatesti (2017) for an argument that even such antisocial people could be construed as having a subject-based reason to play (or to be forced to play) the cooperative game.

4.3 Concluding remarks and possible objections

The goal of this chapter was to further explore in what way subject-based or response-dependentist theories of reasons have the resources to explain why idealization has a natural place inside those accounts. This included the discussion of how such theories could justify evaluations of ends or values that go beyond evaluations of means for satisfying immediate goals. Moreover, this also involved a discussion of the role of idealization in resolving conflicts among individuals and explaining how in that context a pattern of normative reasons that go beyond individual level could emerge.

Yet, it could still be objected that the account of reasons from this chapter and the previous one is too revisionary and that the present discussion does not justify the claim that the reasons that matter are those that are in some sense dependent on our cognitive and affective capacities. In particular, the objectivist could insist that what matters is an objective thing that reflects the intrinsically valuable, desirable, or reason-providing properties of states of affairs. Even in the case of color vision, there are authors who try to save 'commonsensical' mind-independent realism (see, e.g. Tye 2002 ch. 7). Thus, 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.

I take these objections as a cue for the topic of the next chapter. In the following chapter, I will introduce broader naturalistic considerations supporting the view that some version of the subject-based theory of reasons should be favored over object-based theories. The discussion will be based on arguments that rely on considerations from evolutionary biology and psychology.

5 The ontology of normative reasons from an evolutionary perspective

5.1 Introduction

The aim of this chapter is to offer an argument that even if we grant that the commonsensical view of normative reasons presupposes mind-independence, the resulting view is not plausible when evaluated from a naturalistic perspective. The position that will be disputed is a robust version of normative realism (e.g. Fitzpatrick 2008; Enoch 2011; Shafer-Landau 2003; Parfit 2011a; 2011b; 2017). This position can be summarized in three conditions:

1. Normative judgments 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 (see, e.g. 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 facts and that not everything that we believe about, for instance, 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 (see, e.g. Joyce 2006; Mackie 1977; Olson 2014; for recent discussion, see also Taccolini 2024).

For the purposes of the present chapter, condition 3) plays the most important role because it states that what there is a rea-

son to do is stance/mind- or subject-independent. According to Russ 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" (2003, 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.

I will do this by focusing on a specific version of an evolutionary debunking argument (EDA) against the existence of mind-independent normative facts. The literature on EDAs has grown exponentially over the years, especially since Street (2006) and Richard Joyce (2006) have formulated their influential versions of them. Instead of trying to provide an overview of all the different papers covering the topic (see, e.g. Machuca 2023), I will focus on Parfit's underdiscussed objection to EDAs. The importance and novelty of Parfit's discussion is twofold. Parfit points out that normative realists, such as realists about moral facts, have a specific understanding of the normativity of practical judgments, more specifically those pertaining to morality, which makes their origins hard to explain in a way that would diminish their realist credentials. In addition, he offers more general reasons that *prima facie* indicate that the theory of evolution does not have adequate resources to explain the origins of normative judgments about moral matters. Although I will argue that Parfit's objections are not convincing, discussing them will provide us with a significant opportunity to clarify the empirical underpinnings of some EDAs and to overview recent advances in evolutionary explanations of human normative attitudes together with their relevance for metaethical theorizing.

In the rest of the chapter I proceed as follows. I will first distinguish between two types of EDAs. Then, I will review an evolutionary debunking argument pertaining to show that, given that evolutionary forces have shaped the content of our normative or evaluative judgments, it follows that their truth-conditions cannot be completely subject-independent. To reinforce this argument, in the rest of the chapter, I discuss several arguments offered by Parfit against the cogency of this type of EDA.

5.2 Epistemological and ontological aspects of evolutionary debunking arguments

As mentioned, the evolutionary perspective on normative reasons is most often employed in debunking robustly realist/objectivist positions in metaethics (see, e.g. Joyce 2006, ch. 6; Ruse and Wilson 1986; 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 Michael Ruse and Edward O. Wilson's 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" (1986, 186). 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 epistemologically oriented 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, 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 truths about mind-independent reality. Therefore, we cannot be justified in believing that our evaluative judgments, whose formation and maintenance were 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 their own children.[54] 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

54 The example comes from Street (2006, 115).

of those organisms in the population that have greater fitness.[55] 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 2000, 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 becomes relevant: evolution by natural selection is not a truth-tracking process. What is good for spreading genes in some population or for enhancing the survival and reproductive rates of some organism does not have to reflect true states of affairs in any substantive sense (Stich 1990, 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

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

more professional and romantic success than a person whose beliefs about herself are grounded in facts (see, e.g. von Hippel and Trivers 2011).

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 explains 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 (or normative) 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 formulated as having more direct ontological conclusions.[56] This reading of the evolu-

56 Richard 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

tionary debunking argument is actually endorsed by Ruse and Wilson (see, also Rosenberg 2011, ch. 5):

> 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 and Wilson 1986, 186)

In what follows I will develop a discussion that focuses on this type of ontologically oriented EDA, because it seems to me that considerations based on the relation between evolutionary theory and normativity have direct 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 this is so in the specific case of morality, as the above quote states, and in the more general case of normative practical reasons.

In what follows, I will outline, in broad terms, how a type of EDA attacking the ontological grounds of normative realism might be formulated. The main contention is that if we adopt a naturalistic story about the evolution of our evaluative attitudes, then we have a strong explanatory reason for thinking that normative facts to which our evaluative judgments refer, if there are any, cannot be wholly subject-independent—rather, they must be in some sense a function of our attitudes, responses, development, contingent pasts, etc. (see Chapter 3).

5.3 Evaluative judgments, normative reasons and their evolutionary underpinnings

Following Street, I will construe evaluative attitudes as involving "desires, attitudes of approval and disapproval, unreflective

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.

evaluative tendencies (…), and consciously or unconsciously held evaluative judgements (…)" (2006, 110).[57] From a naturalistic perspective, the main function of evaluative attitudes plausibly involves promoting fitness increasing behaviors and deterring from harmful ones (e.g. Joyce 2006; Street 2006). Given that humans are social beings who depend on cooperation with other individuals, this constrains their adaptive landscape. For instance, humans have a long gestation period. Once children are born, they are completely depended on their parents for a relatively long time compared to other primates. Thus, human parents need to dedicate a significant amount of their time to producing and rearing viable offspring. In addition, as adults, our fitness continues to depend on cooperating with other individuals from our social groups. We build things, exchange goods, establish large-scale economies, we help each other when in need, defend against common threats, share information, etc.

Living in cooperative groups, however, is met with different types of conflict of interest that threaten to diminish the benefits of cooperation (Gaus 2011). Most notably, engaging in selfish behaviors by taking advantage of other people's cooperative behavior without paying the costs of cooperation, leads to social conflicts that disrupt social cohesion and schemes of cooperation. In this context, the standard view is that the evolutionary function of morality and capacities that reinforce prosocial behavior is to reap the benefits of cooperation (see, e.g. Haidt 2007; Kitcher

57 The main reason for talking about evaluative attitudes more generally instead of just moral judgments, for instance, is because it is not easy to delineate the moral domain or to determine what would count as a moral judgment as opposed to a normative judgment of another kind (see Sackris and Rosenberg Larsen 2023). This makes it difficult to speculate whether people developed adaptions for morality (see, e.g. Levy and Levy 2018; Pölzler 2018; Cline 2015). Thus, in what follows, I will not presuppose that the moral domain has clear boundaries or that people have specific adaptations for morality (such as a moral sense). In discussing normative realism, I will have in mind normative phenomena that involve evaluative attitudes towards general social affairs and our well-being.

2011; Krebs 2011). In fact, having evaluative attitudes in the form of moral judgments and reactive attitudes plays a significant role in alleviating social conflicts by rewarding mutually beneficial prosocial relationships and punishing disruptive and antisocial behaviors.

On this backdrop, consider the following examples of intuitive judgments whose acceptance could plausibly be explained in evolutionary terms:

i. The fact that something would promote one's survival is a reason in favor of it.
ii. The fact that someone has treated one well is a reason to treat that person well in return.
iii. The fact that someone has done one deliberate harm is a reason to shun that person or seek his or her punishment.
iv. The fact that someone is altruistic is a reason to admire, praise, and reward him or her. (Taken from Street 2006, 115)[58]

We can offer explanations of the emergence and retention of these types of judgments in terms of different evolutionary mechanisms without presupposing that they refer to mind-independent normative facts.[59] The explanation of (i) seems straightforward. It is plausible that if we care about our survival, then caring about the means that enhance our survival will be beneficial for surviving and eventually reproducing. Thus, it is expected that through phylogenetic and ontogenetic development our normative judgments will reflect and be shaped by our more primitive dispositions regarding the preservation of our own

58 Judgments (i)-(iv) instantiate general types. (i) captures a general set of judgments that characterize prudential considerations. (ii)-(iv) characterize a wide set of judgments regarding social morality.

59 Although EDAs presuppose that normative judgments of the type (i)-(iv) can be given an adaptationist explanation, there is no presupposition that these judgments are universally shared or endorsed. Natural selection can maintain variability in the expression of a trait depending on the environment in which an organism is placed and its life history (Stearns 2004, see, also sec. 5.5 below).

lives. Explaining judgments of the type (ii)-(iv) requires recourse to explanations of altruistic behavior.

There are at least five recognized mechanisms by which altruistic behavior could have evolved—including kin selection, direct reciprocity, indirect reciprocity, network reciprocity, and group selection (Nowak 2006). Here, I will focus on the mechanisms of kin selection and reciprocity as they are sufficient to show how we might have come to adopt judgments (i)-(iv).

According to kin selection theory, natural selection maximizes inclusive fitness which involves the ability of an organism to pass on genes by direct reproduction or through helping close relatives with whom we share our genes (Nowak 2006). This theory can explain limited forms of altruism, such as those exhibited between close relatives. For instance, we share around 50% of genes with our children, but we also share around 25% of genes with children of our siblings. Thus, it is expected that we will display more altruistic behavior towards our closer kin, which would decline as our genetic similarity diminishes.

Altruistic behavior towards non-kin can be explained by mechanisms of direct and indirect reciprocity. According to the theory of direct reciprocal altruism, in cooperative interactions one organism—the actor—temporarily incurs fitness costs to itself but increases 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 (i.e. reproduction and/or survival), from living in cooperative groups they have an incentive to endorse cooperative or altruistic behaviors and to punish or shun those who are not altruistic. This would be the tit-for-tat strategy (Axelrod 1984). For example, I help my neighbor harvest her field and in return expect her to help me harvest my field. If the neighbor does not return the favor, I may engage in punitive behavior, such as refusing to help her on the next occasion or more directly acting to reduce her fitness prospects. Therefore,

direct reciprocal altruism can plausibly explain the intuitive appeal of judgments (ii) and (iii).

Indirect reciprocity, which is largely based on the concept of reputation, can explain how we might have evolved capacities for intrinsically valuing things as stated in (iv). Helping someone establishes a good reputation, which will be rewarded by others. When deciding how to act, we consider 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 (see Nowak 2006, 1561). By being helpful across various situations and towards different people, one can build one's reputation in ways that can compensate for considerable costs incurred by such 'altruistic' behavior. In this regard, some experimental studies have shown that helpful people get a positive reputation and receive more benefits in return (see, e.g. Wedekind and Milinski 2000).

As mentioned, many evolutionary mechanisms may explain our normative dispositions, and subsequently our evaluative judgments. Here the relevant point is that the evolutionary theory seems to have the resources to explain our normative judgments. In the context of an ontological debunking type of EDA, this is important for the following reason. If normative realism is true, then our evaluative judgments purport to track an independent realm of normative truths. The theory debunking type of EDA puts pressure on the consequent of this conditional (see, e.g. Hopster 2018; Street 2006, § 6). As we just saw, evolutionary theory can explain the emergence of normative judgments of the type (i)-(iv) without appealing to normative facts that exist independently of and prior to agents' responses or attitudes. Moreover, the evolutionary explanation tells us that even if there were no independent normative truths (or if they were completely different), it would still be adaptive to acquire those evaluative attitudes because they would promote survival and cooperation (Ruse 1986). Thus, given that the existence of stance-independent

moral truths does not play a role in explaining why we endorse familiar sorts of normative judgments, their existence seems to be explanatory and "metaphysically redundant" (Hopster 2018, 7; see, also Harman 1977). This gives us an important reason to deny that such facts exist and to question the veracity of the theory that implies their existence.

This form of EDA fundamentally assumes that evolution significantly influenced the substance of our normative judgments. Nevertheless, Parfit (2011b) presents interesting counterarguments to this assertion. He not only contends that EDAs do not pose a threat to metaethical realism but also challenges the validity of their empirical premise.

Parfit (2011b, ch. 33; see, also 2017) offers three types of considerations against EDAs. First, that the evolutionary theory cannot explain the existence of normative beliefs or why having them would be evolutionarily advantageous. Second, the evolutionary theory cannot explain the pervasiveness of normative beliefs with particular contents (e.g., why we endorse the Golden rule). Third, the evolutionary theory does not offer adequate explanations—especially when there are alternative, non-evolutionary based explanations of our normative beliefs. In the next three sections, I will discuss these objections in turn.

5.4 Normative beliefs from an evolutionary perspective

According to Parfit's first objection, natural selection cannot explain the emergence of normative beliefs, thus EDAs cannot be used to debunk our knowledge of mind-independent normative truths. This objection can be formulated as follows:

1) It was evolutionarily advantageous to be *motivated* to avoid painful stimuli.
2) However, it was not advantageous to acquire the further *belief* that there is a reason to avoid painful stimuli.
3) Thus, since the belief that pain is bad was not advantageous, it is unlikely that we would have formed it if it did not refer

to mind-independent normative facts (Parfit 2011b, 2:529).[60] The main thrust of the argument lies in premiss 2). Indeed, if it was not evolutionarily advantageous to acquire a normative belief, then we cannot explain its existence in evolutionary terms. However, it is not quite clear how to interpret premiss 2). I offer two interpretations and argue that according to either of them, Parfit's argument is not persuasive.

One reading might be that it is not clear why or how we would adopt normative beliefs with specific content because evolutionary processes *cannot* shape their content. It might be claimed that normative beliefs are formed based on individual and social learning, rational reflection, and reasoning.

This interpretation of premiss 2) does not make the argument convincing. Evolutionary processes do not shape and form the contents of normative judgments without the mediation of other processes (Street 2006, 120; for discussion, see also Mogensen 2016). The standard view is that evolutionary forces have partly shaped our intuitive, emotional, and automatic processes that, through ontogenetic development, affect the contents of our more reflective normative judgments (Haidt 2001; Krebs 2005). In social and moral psychology, it is widely accepted that intuitive reactions have primacy over more reflective and conscious processes (Haidt 2007). The idea is that intuitive reactions such as emotional, affective, and other unconscious processes in normal cases cause or in some other way influence the formation of our more conscious and reflective judgments (see, e.g. Haidt 2001; Nichols 2002; see, also Braddock 2016, sec. 7). Let me mention a couple of examples that illustrate how intuitive reactions influence more reflective normative judgments.

Studies show that our negative intuitive reactions towards sex between close siblings account for our tendency to judge that

60 Arnon Levy and Yair Levy (2018, 12) express similar, albeit different, skepticism that currently available evolutionary game-theoretic models of reciprocal altruism can explain psychological altruism.

sex between close siblings is wrong even when it is consensual and there is no prospect of harm (see Haidt 2001). Other studies indicate that our core affective reactions account for why we adopt certain etiquette norms. For instance, Shaun Nichols (2002; cf. May 2014) argues that our disgust reactions towards spitting at the dining table explain why etiquette norms that forbid spitting during eating in the Western societies survived to this today. Similarly, exposure to other people's suffering normally activates negative affective responses in us. Some studies suggest that this predisposition affects our acceptance of norms that forbid harming other people (Blair 1995). Thus, according to the principle of the primacy of the intuitive, it would be expected that our negative responses to painful stimuli have influenced our judgments concerning their badness.

Parfit appears to accept the possibility that the adaptive disposition to avoid painful stimuli somehow "*led* later humans to believe that we have this reason" to avoid painful stimuli (Parfit 2011b, 2:529). Thus, maybe Parfit had some other reading of premiss 2) in mind.

Another interpretation could be that from an evolutionary point of view it is not clear why we would have normative beliefs at all. It makes sense to think that natural selection would shape our motivations and dispositions to act, but it is not entirely clear why it would be, on top of that, evolutionarily advantageous to have *further* beliefs about what we have normative reasons to do.

It might seem that this objection can be easily answered. From an evolutionary point of view, it is plausible to assume that the main function of evaluative judgments is motivational—to reinforce and regulate behavior (Gibbard 1990; Joyce 2006). A general idea is that to benefit from living in cooperative societies, which is typical of human populations throughout evolutionary history, individuals must, among other things, acquire capacities for overcoming temptations to pursue their selfish interests. Thus, they have to develop decision-making capacities that will

be responsive to their long-term interests and interests of other people, and will enable them to act based on these considerations. In this context, the capacity for normative beliefs/judgments can be construed as a psychological (regulative) solution to these selective pressures (see, also Krebs 2005).

However, Parfit may not accept this view of normative judgment, as he appears to construe it as purely representational. In another place he writes:

> when realists appeal to facts about what is normatively necessary, or about what we must do in the decisive-reason-implying sense, these people do not thereby explain how we are *motivated* to act in these ways. That is an objection to normative realism if (…) we assume that normativity is, or consists in, some kind of actual or hypothetical motivating force. But realists reject that assumption. (…) On this view (…) normativity is wholly different from, and does not include, motivating force. (Parfit 2011b, 2:421)

If Parfit holds that normative beliefs lack inherent motivating force, it is reasonable to question whether natural selection would favor them. Hence, the existence of this capacity requires a non-evolutionary explanation.

Several reasons suggest that this interpretation of the argument does not make it compelling. First, the assumption that the function of moral judgments is exhausted by their representational aspect seems to be incorrect. Moral judgments are often accompanied by decisions based on them (Bartels et al. 2014). A plausible explanation of this fact is that representation is not the sole function of normative beliefs; rather, part of their function is to regulate action (Gibbard 1990; Korsgaard 2011).[61]

61 Here I want to remain neutral concerning the motivational internalism/externalism distinction regarding moral or normative judgments. The assertion is not that motivational internalism is preferable from an evolutionary standpoint. Instead, I propose a weaker claim: evolution would favor evaluative capacities capable of causally influencing motivational capacities. In this case, it remains open whether we should endorse motivational internalism or externalism because externalists could also acknowledge that normative judgments may, as a matter of contingent fact, in normal cases (that exclude, for instance, severely antisocial individuals and other non-typical cases) play a causal role

Second, suppose that this is not a problem and that normative beliefs are motivationally epiphenomenal—they in some sense accompany motivation but their main function is not to motivate. Still, this would not mean that evolutionary processes could not influence them *indirectly*. For instance, even if we grant that there is no logical or functional relation between normative beliefs and motivation, still there might be a *causal* relation between them. As already mentioned, studies in moral and social psychology indicate that different types of normative beliefs are causally influenced by affective and intuitive states that are plausibly shaped by evolutionary and developmental processes (Haidt 2001; Nichols 2002). Accordingly, strictly speaking, Parfit could be right that it would not be advantageous to believe that there is a reason to avoid painful stimuli. The capacity for normative belief might have come about by random processes or it could have been a further application of normal representational capacities to the normative domain. Nonetheless, that would not show that there is no regular causal relationship between normative beliefs and motivational states that were influenced by the evolutionary processes and, therefore, that we would not adopt them regardless of the mind-independent normative structure of the world.

Third, if at least part of the function of normative beliefs was not to motivate action, then, contra Parfit, it would be rather mysterious why we have the capacity to produce them. In Parfit's defense, it might be responded that if this is mysterious, then it is no more mysterious than our capacity to think and have true beliefs about modal, logical, and mathematical facts (see Parfit 2011a, 1:489–90). Given that it is plausible to think of mathematical and logical beliefs as representing the non-empirical realm of mind-independent abstract objects, then it should not be too problematic to think that normative judgments refer to the non-empirical mind-independent realm of normative facts.

in reliably producing action.

However, this "companions in guilt strategy" does not help, because it only reinforces the mystery in the case of normative judgments. The question is why would we have the capacity for producing true beliefs about what we have a reason to do, whose function is not related to anything that we actually do? (see Korsgaard 2011) Similar mystery does not arise in the case of mathematical or logical beliefs insofar as we do not think of them as being about inherently *normative* facts.[62] Thus, on pain of changing the topic, modeling our thinking about normative judgments on mathematical and logical judgments will not make the mystery disappear. However, by denying that the function of normative beliefs is *purely* representational the mystery disappears. It seems clear that it is advantageous to have normative beliefs because the capacity to produce them provides a solution to problems of cooperation and behavior regulation, such as feeding, surviving, mating, reproducing, overcoming temptations, furthering long-term interests, coordinating actions, and so on.

It could be replied to the last objection that although normative beliefs do not necessarily motivate, Parfit could still accept that there is a *causal* link to motivation. It is consistent with his view that there is an evolved mechanism transforming some cognitive representations into action-guiding principles or motivations. For instance, there could be a rule of transformation according to which judging that there is a sufficient reason to Φ, *ceteris paribus,* causes one to form the intention to Φ. We can call this rule of transformation the *enkratic disposition* (see Broome 2013, 13).

This response will not support Parfit's case, however. The enkratic disposition could be favored by natural selection only

62 Parfit cannot think of mathematical and logical judgments as inherently normative, in the sense of being about normative reasons. Otherwise, his companions-in-guilt strategy might be considered question-begging, as one could argue that if mathematical or logical judgments are normative, their role would be to regulate how we think, not merely describe a domain of abstract objects (for discussion, see, e.g. Field 2009; Smokrović 2018).

if evaluative judgments are such that, at some point in human history, they reinforced fitness-benefitting behavior. Thus, unless one of the primary functions of evaluative judgments were motivational, it would be unlikely that the enkratic disposition would have evolved. This poses a problem for Parfit's view, because if one of the essential functions of evaluative judgments is to motivate fitness-benefitting behavior, then it cannot be the case that what agents experience as counting in favor of and consequently judge that they have 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 fitness. It follows that a normative realist who thinks that the primary function of evaluative judgments is to objectively reflect mind-independent normative reality cannot explain the existence of the mechanism that transforms those judgments into dispositions to act. Therefore, they would not be able to 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.

To sum up, Parfit's first argument is not compelling because it either falsely presupposes that normative beliefs cannot be substantially influenced by evolutionary processes or it needlessly makes our capacity for normative belief mysterious.

5.5 The argument from the Golden Rule

Parfit's second objection is that natural selection cannot explain particular normative beliefs, such as our acceptance of the Golden Rule, that promises ought to be kept, and that everyone's well-being matters equally.[63] Since Parfit emphasizes the Golden Rule, I will call this objection *the argument from the Golden Rule*. It can

63 Levy and Levy (2018, 10) raise a similar issue. Moreover, Michael Huemer (2016) argues that natural selection cannot explain the world-wide convergence towards what he calls liberal values. For a detailed discussion of Huemer's paper, see Hopster (2020).

be formulated as follows (see Parfit 2011b, 2:536–37):

a) Natural selection can explain our acceptance of normative beliefs that enhance fitness; For instance, it explains how some organisms became *reciprocal* altruists (see, e.g. Trivers 1971).
b) "The Golden Rule, in contrast, tells us to be suckers, who benefit everyone, including cheats" (Parfit 2011b, 2:537).

Here the idea is that the endorsement of the Golden Rule reduces biological fitness because it encourages extreme forms of altruism.

c) Thus, natural selection cannot explain why many people endorse the Golden Rule.

We might be suspicious of the claim that people who accept some standard version of the Golden Rule, such as *One should treat others as one would like others to treat oneself*, understand it as saying that they should be *suckers* in social relations. By "sucker", Parfit presumably means individuals who display extreme altruistic behaviors towards everyone, including those who would take advantage of them without hesitation. One could argue that if people really understood the Golden Rule as demanding pure altruism (without expecting reciprocity), then probably not many people would accept it as a norm of behavior. However, Parfit has a quick retort to this type of objection:

> Natural selection might explain why, of those who have accepted the Golden Rule, most have often failed to do what this rule requires. But we are discussing explanations of our normative beliefs, not our motivation to act on these beliefs. (Parfit 2011b, 2:537)

Contrary to Parfit's suggestion, it is not evident that a common understanding of the Golden Rule necessitates unconditional altruism. In the entry on the Golden Rule in the Internet Encyclopedia of Philosophy, it is described as follows:

> The rule is distinguished from highly supererogatory rationales commonly confused with it—loving thy neighbor as thyself,

> turning the other cheek, and aiding the poor, homeless and afflicted. Like agape or unconditional love, these precepts demand much more altruism of us, and are much more liable to utopianism. The golden rule urges more feasible other-directedness and egalitarianism in our outlook. (Pukka 2023)

In this regard, Parfit may be imposing a reading of the Golden Rule that is too strong. However, let us suppose that he is not, and that people typically understand the Golden Rule as demanding strong forms of altruism. What would be the 'naturalistically friendly' explanation of why people accept such a rule?

In a series of articles, Nicholas Baumard and colleagues (e.g. Baumard and Boyer 2013; Baumard and Chevallier 2015; Baumard et al. 2015) have proposed a plausible explanation within the Life History Theory for the increasing spread of the Golden Rule and similar prosocial norms between 500 and 300 BCE. They suggest that these norms emerged as socially adaptive in affluent societies, fostering the development of moralizing religions such as Buddhism, Hinduism, Judaism, Stoicism, and later, Christianity.

To comprehend their argument, familiarity with the necessary background of Life History Theory (LHT) is essential. LHT, in general, elucidates how evolutionary adaptations to specific ecological niches result in diversity among life histories of various species and individuals within the same species (for overviews, see Stearns 2004; Međedović 2023). Trade-offs arise between life history (LH) traits when they differentially impact fitness. One key outcome of this research is the trade-off between time and energy investments an organism makes over its lifespan to optimize fitness given environmental challenges. For example, longevity and the onset of reproductive efforts are negatively correlated, meaning that selection for longer lifespan favors delayed reproduction (and consequently a lower number of offspring). Fruit flies provide a concrete illustration; experiments demonstrate they postpone reproduction with increased life expectancy

and *vice versa* (Stearns 2004).

Pertinent to our discussion is the correlation and continuum of energy or resource trade-offs. Extremes on this continuum cluster traits that are selected together by natural selection, forming what is known as 'slow' and 'fast' life-history (LH) strategies. Slow LH strategies entail extended longevity, delayed reproduction, greater parental investment, and fewer offspring. Fast LH strategies involve a shorter lifespan, early sexual activity, reduced offspring investment, and increased reproduction (see **Figure 3**).

It is important to note that fast strategies are adaptive in harsh environments. These are unpredictable and uncertain environments with high mortality rates. In these types of circumstances, it pays-off, in evolutionary terms, to "rely on strategies focused on smaller, but more immediate and more certain benefits" (Baumard and Chevallier 2015, 2). In other words, in harsh environments, we can expect that organisms would, on average, develop, start to reproduce, and die earlier. Moreover, as a consequence of early reproduction, they tend to have more progeny. In human terms, these are the circumstances in which it becomes adaptive to act more impulsively, have lower trust in others, and generally expect that others will not be overly cooperative. Alternatively, slow strategies are adaptive in safe environments where "individuals can afford to pursue larger, but less immediate and less certain benefits" (Baumard and Chevallier 2015, 2). In other words, these are predictable, affluent, and safe environments where people tend to invest more in partners, live longer lives, and generally, it pays-off to cooperate more strongly.

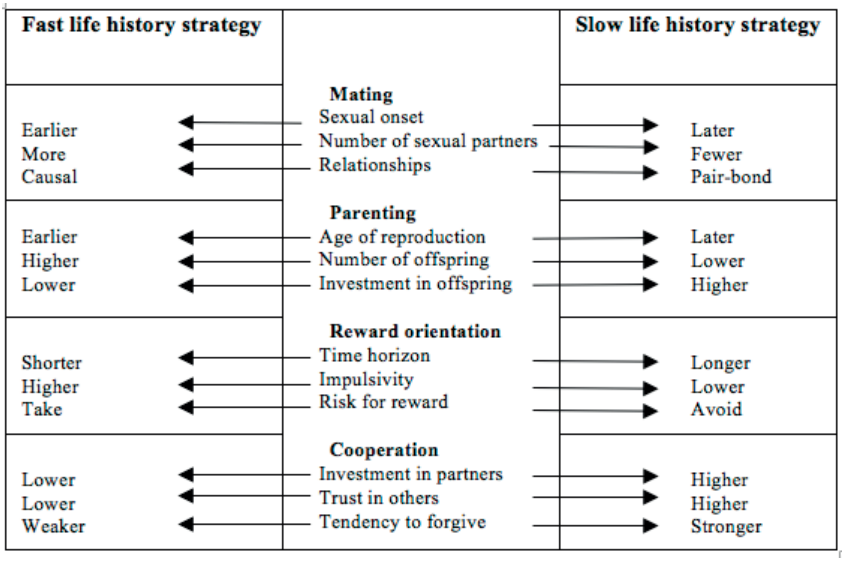

Figure 3: Life-history tradeoffs – fast LHS vs. slow LHS: In the middle column are domains responsive to adaptive trade-offs, with behaviors on the left adapted to poor and unpredictable environments and those on the right associated with rich and predictable environments (adapted from Baumard and Chevallier 2015).

A plausible reason for the emergence of moralizing religions in the first millennium BC is the accumulated wealth in human societies and the ecological niche created by those rich societies. The increase in wealth during that period is typically gauged by citizen calorie intake, population growth, and urban area sizes (Baumard et al. 2015). The guiding idea is that moralizing religions developed in circumstances where the accumulation of wealth fostered an environment favoring slower life history strategies.

Applied to normative beliefs, the basic idea is that such beliefs have been stabilized as reinforcers of particular strategies that are evolutionary adaptive in affluent and predictable ecological niches (Baumard and Chevallier 2015). Evolutionary

approaches to morality tell us that humans have evolved cognitive and affective mechanisms whose function is to encourage fitness-enhancing cooperative behaviors (Haidt 2007; Krebs 2005; see, also Cline 2015). For instance, studies indicate that infants already come pre-tuned with rudimentary capacities for social evaluation based on harm and fairness (Hamlin, Wynne, and Bloom 2007; Smith, Blake, and Harris 2013). Combined with LHT, we can explain why extreme prosocial attitudes and judgments might prevail in certain circumstances. To reiterate, slower LH strategies support a life history in which cooperation is beneficial, long-term planning is adaptive and consequently, norms that reinforce such behaviors are adaptive as well. Normative beliefs such as the Golden Rule can be seen as reflective elaborations of attitudes that express and reinforce our pre-tuned prosocial responses (Baumard and Chevallier 2015; Haidt 2007). The Golden Rule prescribes to individuals to treat others as they would like to be treated which plausibly includes treating others as equals and trusting them in social relations. As such it reinforces our predispositions towards prosocial behavior that is adaptive in affluent and predictable ecological niches (see the right-hand side in **Figure 3**). Thus, it makes sense that the acceptance of norms such as the Golden Rule has emerged in ecological and cultural niches in which it is adaptive to play slower LH strategies.

LHT can also explain the *variability* in norms and their contents across cultures and time periods. For instance, if people live in uncertain environments characterized by scarce resources, then it is expected that fast life strategies become adaptive which would be reflected in behaviors and norms characterized by lower trust in others, greater investment in reproductive efforts, reduced investment in quality upbringing, and adoption of less forgiving "eye for an eye" conceptions of justice (see the left-hand side in **Figure 3**). These claims receive confirmation from studies on the economic behavior of children and adults from different

socioeconomic backgrounds. More specifically, studies show that people from lower socioeconomic environments, regardless of the cultural background, exhibit less altruistic behavior and have less positive attitudes towards prosocial behavior compared to people living in higher socioeconomic environments (Wilson, O'Brien, and Sesma 2009; Chen, Zhu, and Chen 2013; Nettle, Colléony, and Cockerill 2011).

Contemporary evolutionary theory, when combined with Life History Theory, not only provides general explanations for the evolution of altruistic behavior but also offers resources to explain the evolution and spread of specific normative beliefs, such as those endorsing strong altruistic behavior, within human populations. Furthermore, it elucidates how variability in the acceptance of normative beliefs is maintained based on the environment and the individual's specific life history.

5.6 Do cognitive explanations of normative beliefs override evolutionary explanations?

Parfit's third objection posits that, in many cases, it is more plausible to explain our endorsement of certain normative judgments in terms of our ability to respond "to their intrinsic credibility or our reasons to have them" (Parfit 2011b, 2:535). He supports this reasoning with at least two objections to the idea that evolutionary theory provides good explanations for our normative beliefs. One is that an evolutionary approach to explaining normative beliefs gives false empirical predictions. The other is that in the case of normative judgments there are methodological reasons why evolutionary explanations are often not adequate. I will start with the first objection.

Parfit claims that evolutionary theory has some false empirical predictions regarding the content of our normative beliefs. He argues as follows:

> [i]f our moral beliefs were mostly produced by evolutionary forces, we would expect people to believe that they have a duty to

> have and raise as many children as they can, and that deciding not to have children would be wrong. But this is not what most people have believed. (…) If our normative beliefs were selected to maximize the number of our descendants, and of other people who have our genes, these various facts would be hard to explain. (Parfit 2011b, 2:534–35)

In the current context, the objection implies that if evolutionary explanations yield inaccurate predictions about the normative beliefs we endorse, they likely should not be employed to explain our actual beliefs.

However, this objection rests on a mistaken assumption. The theory of evolution does not predict a lack of variation in beliefs about our duties to have children or advocating for having as many children as possible. Kin selection theory already informs us that an individual will act to maximize their inclusive fitness, which does not necessarily entail having their own children. Inclusive fitness can be maximized by aiding close relatives in successfully raising their children. Thus, there is no expectation that people, without exception, will feel the need to have their own children, biasing their beliefs regarding obligations to have children.

Furthermore, LHT explains the expected variation in human normative judgments concerning our duties toward having children. Life-history trade-offs elucidate why, paradoxically, individuals in less affluent countries tend to have more children than those in more affluent societies (for review, see Lawson and Borgerhoff Mulder 2016). In resource-poor environments with lower life expectancy, there is a selection for an earlier onset of reproductive efforts, leading to a higher number of offspring but reduced energetic resources for individual child-rearing. Conversely, in resource-rich environments, a later onset of reproductive efforts allows more resources for individual growth and maturation, as well as quality investment in child-rearing. Such affluent environments can foster diverse beliefs about our duties toward having children. Given the tendency to invest less in re-

productive efforts, it is expected that adopting a belief in having as many children as possible would be less likely.

This leads us to another general objection to using the theory of evolution to explain normative judgments. Parfit proposes a quasi-procedure for determining when to apply an evolutionary explanation to normative beliefs and when to seek an alternative explanation. He writes as follows:

> We can often imagine plausible evolutionary explanations for either two conflicting normative beliefs. This fact counts against both these explanations. Things are different when we consider many biological facts. When such facts raise a problem for evolutionary theory, as is true, for example, of the origin of sexual reproduction, it *may* be enough if we can imagine some fairly plausible evolutionary explanation. We have strong reasons to believe that such facts have some such explanation. No such claim applies to most of our normative beliefs. Since these beliefs can be plausibly explained in other ways, it is not enough to suggest how these beliefs might have been produced by evolutionary forces. (Parfit 2011b, 2:536)

Parfit suggests that if conflicting normative beliefs can be explained through alternative means in addition to evolutionary explanations, reliance on the latter should be reconsidered. For example, we could devise an evolutionary explanation of why raping and committing adultery are believed to be wrong. Had we believed that men ought to rape women and commit adultery, we could have explained this counterfactual situation in evolutionary terms, such as involving an alternative reproductive strategy (Parfit 2011b, 2:535). However, Parfit contends that in such cases, there is an available alternative explanation for why we accept normative judgments, namely their "intrinsic credibility" (intuitive appeal) and our responsiveness to reasons for holding them (Parfit 2011b, 2:535; see, also Huemer 2016). According to Parfit's procedure, in such cases where alternative explanations exist, we should not rely on evolutionary explanations, which leads him to think that it is likely that evolutionary processes did not produce our acceptance of many normative judgments. Thus, their realist

credentials cannot be undermined by EDAs.

Parfit's argument is not persuasive for the following reasons. First, it is one of the tasks of the theory of evolution to explain variation in biological traits as a response to differential developmental and environmental influences (Stearns 2004; Međedović 2023). Thus, it is not clear why the capacity of the theory of evaluation to explain actual or counterfactual variability in human normative beliefs would count against using evolutionary explanations in this context. Second, proposing that cognitive abilities, alongside factors like upbringing and cultural influences, explains the formation of normative beliefs does not diminish the plausibility of the claim that evolutionary processes influenced those beliefs. This becomes apparent when considering the interrelation between proximal and distal explanations. Exploring this topic will also be instructive for discussing implications of the interrelation between proximal and distal mechanisms for normative realism in metaethics.

Evolutionary accounts offer distal explanations for the emergence of traits and their functions, addressing why a trait developed during evolutionary history. In contrast, accounts explaining the emergence or propagation of a trait in terms of cognitive capacities rely on proximal mechanisms, such as attention, memory, decision-making, and reasoning capacities. It is essential to note that the space of possibilities for proximal mechanisms is, in relevant respects, constrained by distal evolutionary processes.[64] By "relevant respects" I mean the standard view that evolutionary processes had a significant role in shaping our mechanisms

64 For a classical statement of the distinction between proximal and distal explanations in biology, see Mayr (1961). Andreas Mogensen (2015) also emphasizes the importance of the distinction in the context of EDAs. Surprisingly, he argues that making this distinction should mitigate the force of different EDAs against normative realism. Mogensen puts too much emphasis on the distinctiveness and isolation of these two types of explanation and fails to consider how distal explanations constrain proximal ones, which I explain in the main text. For a critical discussion of Mogensen's paper, see FitzPatrick (2016) and Severini (2016).

for acquiring beliefs of the type (i)-(iv); they shape the inputs on which those belief-forming mechanisms work, they shape their operations, output conditions, and the way in which they operate in concert or isolation from other mechanisms (see, e.g. Street 2006). This is expected from an evolutionary perspective because cognitive capacities underpinning social and prudential decision-making are most centrally related to biological fitness (see Sober and Wilson 1998, 159).

To illustrate the entanglement of proximal and distal mechanisms more vividly, consider the following simple model of how we might have developed a capacity for flexibly producing normative judgments that are responsive to biological fitness. The natural tendency of any well-adapted organism (including people) is to maximize its inclusive fitness (El Mouden et al. 2012).[65] However, organisms typically do not *consciously* maximize their inclusive fitness; rather, they appear to develop capacities for responding to cues reliably associated with fitness. In our case, these proximal mechanisms can be, following Claire El Mouden and colleagues (2012), broadly classified under the terms 'pleasure' and 'pain'. For example, the pleasure of engaging in sexual activity served as a reliable cue to reproductive success. The risk of burns associated with fire functioned as a deterrent, serving as a cue to maintain distance. The evolution of such positive and negative affective/motivational states further grounded other more sophisticated emotions and cognitive abilities. We can imagine that when people became self-conscious, they envisaged those affective states as the most elementary components grounding their actions (Korsgaard 2011). Thus, the evolution of proto-normative states results in content that can be broadly characterized as "Pleasure is good", "Pain is bad", "More plea-

65 Roughly, this means that we act in ways that support the spreading of our genes and those shared by our close relatives. 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 discussion of these issues, see El Mouden et al. (2012).

sure is preferable to less", "Less pain is preferable to more", and so forth. In essence, this process could be interpreted as instilling an intrinsic valuation of pleasure and the acquisition of corresponding normative beliefs. Furthermore, it provides an explanation for our endorsement of such beliefs without assuming that they refer to normative facts independent of response or perspective, and historical context.

However, it might be objected that what has been said so far does not undermine stance-independent normative realism. Realists could offer their own "story" about how we might have developed autonomous reasoning skills that are responsive to mind-independent normative reality. For instance, Parfit claims that "just as cheetahs were selected for their speed, and giraffes for their long necks, we were selected for our ability to respond to reasons" (2011b, 2:528; see, also Street 2006, sec. 6). We might further suppose that by using those reasoning capacities, we structured our environment in a way that enabled us to take control and create selection pressures that were at least *indirectly* responsive to response-independent normative reasons.

However, for this response to be compelling, it is necessary to presuppose something along the following lines. At some point in human history not only have we started to value pleasurable (and other fitness related) states intrinsically, but some people discovered that our valuing them reflects mind-independent normative facts (see, e.g. Enoch 2010). Because they were smart enough, they managed to establish social rules and institutions that reflected these mind-independent facts, and given their social influence, they managed to propagate these and other truths among the rest of the population (see, e.g. Huemer 2016). Moreover, for this interpretation to work, it must be taken that the autonomous cognitive capacities that enabled them to respond to mind-independent reasons also enabled them to cognitively, if not bodily, detach from fitness-relevant biological influences. Otherwise, one might object that the autonomous cognitive ca-

pacities mentioned by Parfit are not genuinely autonomous, as they are shaped by a preceding selective history that aligns with factors linked to fitness. Consequently, there would be no need for a reference to objective normative facts. Indeed, we can imagine how people's motivational sets were detached from their tendency to maximize inclusive fitness. The development of more sophisticated cognitive abilities allowed people to seek and find better and more efficient ways of attaining pleasurable states, which ultimately enabled them to detach from their biological origins. For example, greater cognitive abilities enabled people to invent methods of birth control and to have safe sex without worrying about accidental pregnancies. This invention might have even lowered people's inclusive fitness. Thus, the invention of modern technologies might have changed the ecological or cultural niche together with the selective pressures acting within them so that they started to reflect the mind-independent normative facts.

To give it more scientific credentials, this idea might be fleshed out within the gene-culture coevolution view and the niche construction theory (see, e.g. Laland 2008; Richerson and Boyd 2005). According to these views, 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, 3578). The core of the view is that culture, i.e. social, and individual learning are important sources of genetic evolution, in the sense that culture and our cognitive mechanisms can modify the environment, which in turn modifies the selective pressures that act on genes. A well-known example is the coevolution of lactose absorption and human dairy farming. It is thought 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.

There are at least two interrelated problems with this story. One issue pertains to the idea that we can devise norms and propagate cultural traits that will be unconstrained by biological considerations. In fact, it should be noted that formal analysis of the relation between the evolution of genes and culture has shown that genetic selection limits which cultural items (such as beliefs, behaviors, norms, institutions, etc.) will be favored by natural selection and that evolved human cognitive biases constrain and tend to eliminate cultural traits that are biologically maladaptive (for discussion, see El Mouden et al. 2014). Thus, in the long run, only those cultural traits that are advantageous or neutral concerning inclusive fitness maximization might be expected to survive. These considerations indicate that whatever our cognitive and social abilities can devise, and in turn shape our environments, will be constrained by the space of possibilities allowed by our contingent evolutionary history (for a related point, see Severini 2016, 873–74).

Thus, to make the story plausible, normative realists owe us an explanation of why autonomous proximal explanations of our normative beliefs require the existence of independent normative reality whose possible content is constrained by our evolutionary history. This is a challenging task because it is not clear why we should suppose that independent normative reality resides exactly within a space of possibilities constrained by our contingent evolutionary history. If we do not believe in cosmic coincidences, then we should not expect this independent normative reality to be outlined by blind evolutionary processes (Street 2006; see, also Hopster 2019). However, if there is no reason to suppose that this independent reality resides within a space of possibilities constrained by our contingent evolutionary history, then this notion seems to be explanatorily vacuous when it comes to explaining our evaluating attitudes. This leads us to another more general problem.

The more general issue is related to the non-parsimonious assumption about the postulated independent normative facts. Against Parfit, I have argued that the theory of evolution has resources to explain our capacity for normative attitudes and their contents. Thus, dialectically speaking we are in a situation where it seems we can explain our normative judgments as indirectly shaped and filtered by natural selection. In this context, the mind-independent normative realist introduces additional assumptions for which there is no obvious justification (see Street 2006, sec. 6). The realist wants to say that on top of everything that is described in the simple model, for instance, the fact that a person *takes* something to be intrinsically pleasurable and therefore starts to value it intrinsically means that we need to add further ontological ingredients—namely, that the things that are valued intrinsically (or being judged as valuable) have a *further* property of being mind-independently valuable (or reason-providing). It should be emphasized that here the problem is not whether our normative beliefs are epistemically justified. Even metaethical irrealists or response-dependentists can agree with Parfit (see, e.g. his 2011b, 2:539) that normative beliefs will often be justified by normative facts about reasons we have to endorse them. The question is why we should suppose that these reasons are made true by an independent order of normative facts. The only plausible argument for introducing such additional ontological assumptions would be that we cannot explain the possession of certain normative judgments in evolutionary terms or that we have more plausible alternative ways of explaining them. That is, if we had grounds to suppose that the realm of facts to which our normative judgments pertain was not delineated by contingent evolutionary processes. However, as we have seen with Parfit's examples purporting to show the limits of evolutionary explanations, this is either not true or it disregards the relation between proximal and distal explanations (see, also Hopster 2020; Severini 2016).

Of course, what propels normative realists to think that they have some kind of justification is the *intuition* that certain things have intrinsically valuable or reason-giving properties. Parfit sometimes expresses this intuition dramatically, by claiming that "[i]f there are no [irreducibly normative, reason-involving] truths, nothing matters" (2011b, 2:465; for discussion, see Street 2017). But once we recognize that these intuitions are probably grounded in the same mechanisms that process fitness-related cues (i.e. pleasure and pain), they are undercut as reasons for thinking that what we value intrinsically must also refer to the alleged mind-independent normative reality. The additional claim that those things have actual intrinsic value (or reason-providing response-independent properties) does not strictly play any role in the explanation of how our basic normative attitudes were formed or the fact that we may have evolved capacities that are detached from valuing direct fitness-relevant considerations. The bottom line is that even if the hypothesis that our cognitive abilities, through some historical and conceptual development, took control over our biological nature were plausible, the hypothesis that our normative beliefs now reflect response-independent normative reality would still be superfluous.

5.7 Conclusion

In this chapter, I argued that evolutionary debunking arguments bolster the perspective that a naturalistic understanding of normative reasons implies their probable dependence on cognitive or response-related factors. To additionally support this view, I explored often-overlooked yet significant arguments presented by Parfit (2011b), asserting that evolutionary theory inadequately accounts for the origin and persistence of normative judgments. I argued that Parfit's objections fail in ways that can only be discerned if we reflect more deeply on the resources and methodological commitments of evolutionary theory. Consequently, addressing these objections provided a valuable opportunity to

elucidate the empirical foundations of certain evolutionary debunking arguments and the explanatory capabilities of evolutionary theory in understanding various aspects of our normative judgments and their contents.

6 The emergence of reasons and rationality

6.1 Introduction

The objective of this chapter is to formulate a framework for normative reasons that adheres to the constraints outlined in the preceding discussion of the evolutionary argument but can also accommodate plausible features of different types of normative reasons. Thus, in this chapter I will explore how a naturalistic theory of normative reasons could explain the difference between reasons that we experience as depending on our conative and cognitive make-up and those that we experience as transcending particular occurrent desires, goals, or aims. In Kantian terminology, the former could be called *hypothetical* reasons, while the latter could be called *categorical reasons.* As a first step in this discussion, I construe the difference between these two types of reasons as phenomenological.[66] 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.

I will argue that a subject-based theory of reasons can effectively explain the phenomenological distinction between the two types of reasons. I aim to show this by constructing a naturalistic narrative sketching the emergence and stabilization of reasons through the responses of agents with varying levels of cognitive complexity. In formulating this narrative, I will posit that the concept of rationality serves as the cornerstone for discerning the origins of our practical reasons. Through this exploration, it will become evident that the differentiation between hypothetical and categorical reasons hinges on the specific rational principles we embrace.

66 Understood in this way, we do not prejudge that this difference corresponds to an *ontological* distinction between hypothetical and categorical reasons.

In this chapter, I proceed as follows. Initially, I explicate my understanding of the distinction between hypothetical and categorical reasons. Subsequently, I will examine the interconnection among three pivotal concepts: the faculty of reason, rationality, and substantive reasons. Adopting a perspective wherein the faculty of reason and its operational principles determine our substantive reasons, I justify this stance from a naturalistic standpoint, emphasizing the application of distinct rationality criteria based on agents' levels of cognitive and behavioral complexity. Following this, I will introduce principles for differentiating hypothetical and categorical reasons. The principle of instrumental rationality suffices to explain hypothetical reasons. To discern categorical reasons, postulating more substantive principles becomes necessary. Historically, naturalists have encountered challenges in elucidating how we might adopt principles surpassing the instrumental rationality principle. To address this, I will employ a game-theoretic model elucidating the establishment of primitive semantic relations within a community of agents. I will argue that this model can be extended to serve as a model for the emergence of relations governing categorical reasons.

6.2 Hypothetical and categorical reasons

As previously noted, there appears to be an intuitive distinction between at least two categories of reasons: hypothetical and categorical. Thus, a robust theory of reasons should possess the capacity to delineate between these two types of reasons. Alternatively, if a theory fails to make such a distinction, it should offer an explanation as to why, contrary to initial appearances, this differentiation does not hold.

Hypothetical reasons are commonly understood as reasons that fundamentally rely on agent's desires, broadly interpreted. The term "essentially" in this context denotes that reasons are contingent upon a specific agent's motivational framework and its individual components: if a desire is part of the set, it con-

stitutes a reason to fulfill it; conversely, the absence of a desire implies the absence of such a reason. To illustrate this idea, Jonas Olson provides 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, 118)

Conversely, categorical reasons are typically thought of as not relying contingently on the specific desires of the agent. A paradigmatic illustration of how we conceive categorical reasons is derived from moral requirements. Once more, an example provided by Olson can elucidate this distinction:

> 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, 118–19)

Categorical reasons, exemplified by moral reasons, are supposed to possess a form of *inescapability* that hypothetical reasons lack; it appears that they cannot be simply disregarded by merely los-

ing a desire to adhere to them (Foot 1972; see, also, Ventham 2023). Beyond categoricity and inescapability, certain authors assert that moral reasons, in particular, exhibit an (overriding) *authority*. This implies that when these reasons conflict with other non-moral reasons, they generally take precedence and prevail (see, e.g. Brink 1997).

If one embraces a subject-based theory of reasons, accommodating hypothetical reasons poses no inherent difficulty. In such theories, reasons stem from facts about an agent's desires, goals, and concerns. However, categorical reasons might present a challenge, as they are expected to apply universally, independent of an agent's contingent aims or concerns (Ventham 2023). Nevertheless, I will argue that categorical reasons can be conceptualized as a contingent extension of an agent's hypothetical reasons—essentially, they are subjective reasons on a broader scale. I will further argue that as such categorical reasons emerge through interactions between diverse agents, thus representing hypothetical reasons that arise from a population of agents and are applicable to individuals based on their membership in a specifically structured population.

6.3 Reason, rationality, and substantive reasons

In normative philosophy, a distinction is often made between three fundamental concepts: the faculty of reason, rationality, and substantive reasons (see, e.g. Korsgaard 2011; Schafer 2018). Reason, as a faculty, is commonly understood as an active aspect of the mind endowed with a distinctive authority over our thoughts and actions, distinguishing human cognition. In this framework, rationality can be construed as a collection of principles delineating the appropriate functioning of the faculty of reason. Substantive reasons, on the other hand, encompass specific entities, facts, or states of affairs that favor a particular course of action, constituting the elements to which the faculty of reason responds.

Various authors interpret the relationship between these three concepts divergently. In Chapter 2, we noted that Parfit (2011a; 2011b) and other proponents of the irreducible normativity of normative reasons often highlight substantive reasons, tending to explicate rational capacities in terms of them (see, e.g. Lord 2018; Scanlon 1998; Raz 1975; Rowland 2019). In contrast, other influential authors, such as John Broome (2013), contend that rational requirements and substantive reasons are distinct, with neither being convincingly explained by the other. Yet others, exemplified by Christine Korsgaard (2011; see, e.g., also Smith 2013; Schafer 2015c; 2015b; Way 2017), assert that the faculty of reason serves as the fundamental source of normativity, and the nature of substantive reasons can be elucidated in terms of this faculty. In this context, I align with the latter perspective, the so called, capacity first approach to normative reasons (Schafer 2018).

An important reason for adopting this perspective is that the alternative perspectives, wherein substantive reasons are considered entirely distinct, appear implausible to me. This is because views that seek to explicate rationality in terms of substantive reasons can be ultimately categorized into two types: those asserting that substantive reasons can be clarified in terms of rational requirements, and those positing that rational requirements and substantive reasons are fundamentally distinct entities. 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). If one is disinclined to accept the notion that intuitions about substantive reasons can be construed as intuitions about the demands of rationality, I maintain that this reluctance likely stems from being influenced by intuitions similar to those underlying the Williams' gin-and-tonic example (see, e.g. Broome 2007, 167). The intuitions that many seem to have is not merely

that Mary acts rationally when consuming the petroleum under the belief that it is gin and tonic, but rather that she would act irrationally if she refrained from doing so, even though she lacks any objective reason to ingest the petroleum. If one finds this intuition compelling, it suggests an inclination toward the idea that rationality could mandate actions independent of one's actual reasons.

My reluctance to embrace this perspective is tied to how proponents of this view tend to interpret the concept of reason (for recent discussion, see Fogal and Risberg 2023). For instance, Broome interprets reasons as a specific type of explanation for ought-facts. As an illustration, consider Broome's definition of what he terms *pro toto* reasons: "A *pro toto* reason for N to F is an explanation of why N ought to F" (Broome 2013, 50). In this perspective, a normative reason is conceptualized as a fact that necessitates a particular state of affairs, akin to how natural selection necessitates the occurrence of evolution (see Broome 2013, 48).

This proposal raises a concern as it fails to adequately capture the function of reasons in deliberation and may mislead us into believing that the results of reasoning involve judgments that reference some independently existing ought-facts.[67] For example, when I think that I have a conclusive reason to believe that proposition p, it does not necessarily follow that I automatically come to believe that these reasons offer an explanation for why I ought to believe that p. On the one hand, by recognizing normative reasons to believe in p, I might simply acknowledge that, according to epistemic norm E, I am justified in asserting that p is the case. On the other hand, I could maintain that the premiss-

67 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.

es leading to my conclusion logically entail the conclusion itself, without necessarily believing that these premises inherently dictate that I ought to believe the conclusion in a manner external to the deliberative processes guiding me to that conclusion.

Furthermore, in the most extreme scenario, I might not hold the belief that there is any definitive action or belief that I truly ought to adopt. Nevertheless, even in this extreme circumstance, I could maintain the perspective that there exist superior and inferior reasons for believing certain things, as well as more effective and less effective methods of carrying out tasks. It appears to me that even in such cases a certain level of normativity would persist and necessitate explanation. For instance, even if there were no purely mind-independent normative facts dictating what actions or beliefs are correct, we would still encounter challenges that demand resolution and decisions that require consideration. The conclusions that we would reach would often involve the idea that we *should* do something. Nonetheless, this judgment regarding what we ought to do remains of a practical nature and does not constitute a representation of an objective fact that requires explanation during deliberation. If this judgment were to be characterized as true, its truthfulness would be contingent on something inherent to the process that led to it. Naturally, this process would be denoted as normative, but the normativity involved would align with the kind typically associated with the rationality of deliberation and the tasks we are disposed to undertake. In Korsgaard's words, we naturally come to the view that "if reasons did not exist, we would have to invent them" (2011, 6). We would have to devise reasons to fulfill a practical function in guiding action. From this standpoint, it becomes evident that postulating reasons as theoretical entities within a detached normative realm does not contribute anything substantive to the practical role that reasons play in our cognitive economy.

In addition, Korsgaard's construal of the situation has naturalistic credentials. This approach furnishes us with conceptual tools that can be integrated with notions derived from cognitive and evolutionary sciences. Let me illustrate the general idea. In this perspective, substantive reasons do not emerge as something inherently peculiar or ontologically irreducible. Instead, reasons can be understood as entities that furnish inputs to the faculty of reason, and what they count in favor of or support aligns with the outputs produced by the faculty of reason when it operates effectively. Thus, the emphasis is placed on the faculty of reason and its principles of rational operation. The subsequent discussion will be about how to conceptualize these principles and how they can, in a manner consistent with naturalistic principles, account for the distinction between hypothetical and categorical reasons. Addressing this question can begin by considering the function of the faculty of reason and its guiding principles.

6.3.1 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). Moreover, in the context of engaging in tasks, the concept of rationality appears most aptly suited to situations where an organism is confronted with a 'space of alternatives' from which it can select types or tokens of behaviors (see Bermúdez 2003, 117). The primary task for every living creature is to endure long enough to engage in reproduction. However, depending on the specific task that a creature is undertaking, diverse forms of rationality evolved as essential conditions to effectively carry out the task. José Bermúdez (2003; see, also Kacelnik 2006) helpfully distinguishes between three types of rationality (that characterize three types of faculties of reason) that we can ascribe to creatures.

At the most basic level we find what Bermúdez (2003, 116) calls level 0 rationality. This type of rationality is basic in the sense that it involves the ability to form and learn adaptive re-

sponses 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 (see, e.g. Brembs 2009). According to Bermúdez (2003, 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 this sense, when assessing an organism's level 0 rationality, the inquiry does not revolve around whether any specific action aligns with a particular goal—as genuine decision-making may not be requisite. Instead, the evaluation centers on patterns or programs (algorithms) governing behavior to which the organism is predisposed. These behavioral patterns can be instantiated at the level of genetically encoded hard-wired behavioral procedures, but not exclusively, as they may also encompass domain-general learning systems like classical and operant conditioning.

At this rudimentary level of rationality, even observable in organisms like fruit flies, behavioral dispositions are appraised based on both short-term and long-term criteria. Among the latter, Bermúdez (2003, 118), following Richard Dawkins (1986) and others, incorporates the organism's overarching endeavor to maximize its inclusive fitness. The former criteria encompass the fulfillment of more immediate objectives, such as optimizing energy intake, negotiating trade-offs between exploratory and exploitative efforts during foraging, and balancing specific activities (like mating and evading predators) linked to reproduction and survival. As we will explore further, all subsequent levels of rationality will involve analogous short-term and long-term evaluation criteria.

At the top of the conceptual hierarchy of rationality is what Bermúdez (2003, 123) calls level 2 rationality. This constitutes the fully developed, commonsense concept of rationality, encompassing a sophisticated representational framework, a theory of

mind, and the capability to integrate various mental states in decision-making processes. At this stage, rational evaluations extend to both specific actions (not solely types of behaviors) and the decision-making processes themselves. Positioned between levels 0 and 2 is level 1, distinct from level 2 in that it lacks a sophisticated representational apparatus or decision-making and distinct from level 0 as it permits the application of rational standards to token behaviors or actions. This level of rationality holds significance in the current context as it already involves a recognizable form of normativity. To elucidate this, let us explore how Bermúdez approaches this issue.

The fundamental characteristic of creatures with level 1 rationality is their ability to perceive the world (environment) as segmented into opportunities for action, allowing them to choose alternatives based on their predetermined needs or goals, all without partaking in any substantial or folk-psychologically familiar decision-making. Bermúdez illustrates this concept with 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, 121)

In this context, Bermúdez, drawing on James Gibson (1979), employs the concept of affordances. This concept elucidates a type of immediate perception that can explain behavior without necessitating the assumption that the organism engages in cognitively specified decision-making. The notion of an affordance allows us to recognize that perception involves more than just sensing objective spatial and temporal relations in the environment; instead

> [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, 121)

At this level of rationality, the concept of affordances facilitates the integration of normative reasons with a notion that is compatible with naturalistic principles. It aids in unpacking the responsive aspect of dispositionalist accounts of reasons (see, also, Starzak and Schlicht 2023). While affordances, as possibilities of actions, are objective, the determination of which action possibilities are relevant is still influenced by the abilities, needs, and tasks that an organism has evolved to perform. According to Gibson (see his 1979, 128), affordances are relative to individuals. For instance, a child perceives a tiny chair as sit-on-able, whereas an adult, being too tall for the chair, does not. In this sense, the relevance of the affordances provided by an environment is shaped by the responses the organism is predisposed to make and the advantages it thereby gains.[68]

Specifically, affordances provide a framework for understanding the world as normatively imbued. At the phenomenological level, we perceive things and situations as presenting opportunities for action or, in more familiar terms, as indicative of what counts in favor of taking one course of action as opposed to another. In fact, when addressing the genesis of reason, Korsgaard describes the situation in comparable terms:

68 To draw the analogy that I am attempting to make, it is crucial to emphasize that Gibson does not conceptualize affordances as entirely 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, 129). Thus, affordances can be naturally interpreted as response-dependent properties. This is the sense in which I think the notion of an affordance can be used to illuminate the fact that the world is normatively given to us.

> 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, 10–11)

We observe that the commonplace form of normativity is already evident in level 1 rationality. At this stage, there is not a distinct demarcation between various mental states, such as beliefs and desires; rather, the world appears to be presented to creatures in a more directly organized manner through affordances. In other words, affordances can be construed as providing fundamental normative categories that are presented to us in relation to our needs, preferences, and the tasks we are undertaking. Organisms susceptible to evaluation in terms of level 1 rationality exhibit increased flexibility in behavior, responsiveness to environmental cues, and action selection. 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 the action. This enables organisms to

learn more flexibly and adapt to changing environments, and to avoid the constraints of hardwired behavioral dispositions.

Consistent with this perspective, Bermúdez highlights that level 1 rationality is amenable to evaluation based on both short- and long-term criteria. Once more, long-term criteria pertain to maximizing inclusive fitness, while short-term criteria relate to immediate goals that, in the broader context, should contribute to long-term goals. However, given the heightened flexibility of behavior and adaptability in learning action-potentials at this level, it becomes more feasible to assess specific actions in relation to specific immediate goals. The increased flexibility inherent in level 1 rationality accommodates the possibility of conflicting rational evaluation criteria.

For instance, Bermúdez (2003, 121) highlights that vervet monkeys possess a sophisticated signaling system allowing them to alert one another when a predator is approaching. While having such a signaling system yields long-term fitness advantages for the entire vervet monkey population, this is contingent upon a sufficient number of individuals actively participating in the activity of alerting.[69] Nevertheless, engaging in such a community introduces the possibility of adhering to different criteria of rationality. For instance, a vervet monkey that opts to flee when confronted with a predator, rather than remaining to warn others, might be acting rationally in more immediate terms. However, this behavior may not be considered rational in terms of long-term inclusive fitness, provided that a sufficient number of other monkeys fulfill their roles in the community's warning system.

69 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 their 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.

As already mentioned, at the apex of the hierarchy is level 2 rationality. The most notable distinction at this organizational level is that the organism possesses the capacity to adaptively respond to cues from the environment and the cognitive ability to reflect, consider its representations of the environment, and participate in a comprehensive decision-making process. This marks the level of cognitive sophistication wherein a creature can become cognizant of the underlying reasons for its actions and thoughts, consequently gaining control over them (see, also Dennett 2003, 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, 11). In this regard, Korsgaard continues:

> [O]nce 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, 11)

According to Korsgaard, the source of reason lies in the ability to reflectively contemplate the grounds, reasons, or 'rationales', as Daniel Dennett (2003, 204) would label them, for our actions. A naturalistically conceived hierarchy of cognitive abilities implies that a recognizable form of normativity is already inherent in our perceptions of affordances, and is not necessarily generated at the level of self-reflective conscious reasoning. However, for Korsgaard and others within the Kantian tradition, it appears that reasons are uniquely individuated at the level of 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" (2011, 11). If we interpret this statement as asserting that a prerequisite for something to be considered a reason is for us to endorse it by representing it as a basis for our beliefs or actions, then this would seemingly rule out level 1

rationality and affordances as sources of reasons. The rationale for this is that, as per Bermúdez, level 1 rationality does not necessitate decision-making involving higher-order thought. There are at least two reasons to question the plausibility of Korsgaard's view; one is conceptual, and the other is more empirical.

First, from a conceptual standpoint, Korsgaard's view could give rise to an infinite regress. As Peter Railton (2004; 2009) has argued in a similar context, if we assume that a consideration becomes a reason when we endorse it as a basis for action, the question then arises about what endorsement means in this context. One natural proposal is to interpret it as some form of action, perhaps a (mental) approval on our part. However, when interpreted in this manner, we naturally begin to question whether this act of approval is justified or supported by reasons. If it is not, then it is unclear how that endorsement could render a consideration into a reason. However, if supportive reasons are indeed normative reasons, they too should be endorsed, as rational endorsements transform considerations into normative reasons. Given that the question could be raised again at this juncture, we can see how the infinite regress might unfold.

Alternatively, we could interpret endorsement not as an action but as a form of susceptibility or a feeling that certain grounds count in favor of and lead to a particular response (see Railton 2004, 194–95). However, if we adopt this second interpretation, then we find ourselves in the realm of level 1 rationality. As mentioned, the notion of counting in favor of, at this fundamental level, appears to align well with perceiving affordances. In this specific example, it is linked with having affective or intuitive responses that do not necessarily hinge on our capacity for self-reflective contemplation of the grounds of our thoughts and actions.

The proposition that basic reasons originate from level 1 rationality also aligns well with a naturalistic viewpoint. From an evolutionary standpoint, agents with more intricate deci-

sion-making systems are likely those capable of perceiving affordances and undertaking more sophisticated actions. However, given the assumption that agents of varying complexities exist on a motivational, affective, and cognitive continuum, these more fundamental normative categories and perceptions of the world would likely persist in influencing the decision-making processes of more sophisticated reasoners.

This point is exemplified by the phenomenon of moral dumbfounding (see Haidt 2001). When an average person is challenged to justify the judgment that incest is wrong, they typically search for reasons related to harmful consequences for individuals engaging in incestuous relations. However, even when a psychologist, playing devil's advocate, refutes all reasons pointing to the idea that incest is wrong, people often retain the intuition that incest is wrong.[70] Jonathan Haidt (2001) describes this as a state of being dumbfounded—individuals experience a strong feeling that incest is wrong but struggle to articulate reasons for their judgments. The explanation lies in the fact that, for us, the world is already presented as normatively circumscribed. These intuitions, situated further along the cognitive continuum, can then feed into our more reflective deliberative system, where they may compete with other intuitions or be evaluated in alignment with additional intuitions or reasoning criteria that we adopt.

However, Korsgaard and other Kantians are correct in emphasizing that what distinguishes human agents is their capacity for decision-making, which, as outlined by Bermúdez (2003), underlies level 2 rationality. Full-blown decision-making introduces distinct criteria for evaluating rationality. At the most fundamental level, we encounter familiar criteria for assessing instrumental or procedural rationality. This encompasses acting

70 For instance, psychologists participating in the study may defend a couple engaging in sexual activity by arguing that the intercourse would occur only once, the partners would use protection, everything is consensual, and they love each other, among other justifications.

on the basis of reasons or grounds that are explicitly represented, such as when we act based on an evaluation of the various consequences that potential courses of action might lead to. This involves assigning desirability values to potential action-consequences, along with holding instrumental beliefs regarding the likelihood of achieving various goals in line with their values. Decision-making can encompass choosing based on different criteria, not solely those dependent on the consequences of a particular action. For example, deontologists highlight that we can make choices in accordance with the principles we adopt (see, e.g. Gaus 2011). This may involve acting on an intention that can be appropriately universalized or is acceptable to all parties involved in a decision-making process, among other criteria.

The decision-making and its components in level 2 rationality are subject to markedly different criteria compared to levels 0 and 1 rationality. The organism's capacity to create detached representations of its environment and its values enables a more internally based assessment of rationality. Once again, a distinction can be made between more distal and proximal criteria of rationality. Distal criteria pertain to fitness considerations, while the proximal criteria become even more nuanced. For instance, at this level, we can evaluate specific mental states and their contents, regardless of how well they correspond to reality. This introduces a higher potential for conflicting judgments about the rationality of an agent. This explains the familiar phenomenon that a person can be rational in her beliefs and actions even if the action or belief does not meet some externally imposed criteria (such as aligning with reality, offering fitness benefits, effectively fulfilling an intended goal, etc.). For example, Mary may be rational in drinking from a glass full of petroleum, even if that action does not align with her desires or fulfill her other aims. The reason she might be considered rational in drinking from the glass is because she believes that the glass contains gin and tonic.

The potential for conflicting criteria allows us to differentiate between reasons that are normatively given due to their individuation at level 1 rationality and those originating from more sophisticated decision-making processes involving detached representations and environmental evaluations. The relationship between the two levels can be conceptualized as follows: basic affordances perceived as external, along with other internally based instincts, initially constrain our decision-making processes at a more cognitive level. What we perceive phenomenologically as counting in favor of something will determine the values we seek to pursue at a more cognitive level of decision-making. As a first approximation, we might state that level 2 rationality will be evaluated based on how effectively it satisfies the goals set at level 1 rationality. Naturally, however, our ability to contemplate our representations and their meanings, and to exert control over their grounds, will empower us to alter the evaluations stemming from a more primitive level.

To illustrate this point, consider the phenomenon of implicit biases that many people may hold against individuals from other races. However, through the top-down influence of our more cognitively sophisticated decision-making processes, we can suppress and even eliminate these biases (see, e.g. Kennett and Fine 2009). However, the key point I want to emphasize is that instead of assuming that top-down processes control everything and that level 2 rationality criteria should dominate all others, we should consider an interactive loop between levels. The idea is that primitive normative representations originate from more primitive decision-making processes and needs.[71] These primitive normative representations then feed into the more cognitively based representational system, which, through a feedback loop,

71 For a discussion of the notion of need, see David Copp (1995, ch. 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.

can influence these more primitive processes.[72] In this sense, the faculty of reason encompasses both more evolutionary and cognitively basic processes, as well as more cognitively and reflectively sophisticated ones. In this conceptualization, substantive reasons are considerations originating from different levels of decision-making that interact, compete, and serve as grounds for more reflectively laden decision-making processes.

The crucial point to emphasize is that all three levels of rationality and the reasons for action they determine are defined by external criteria. In other words, the criteria of rationality are established in reference to the assumed task that the organism is performing and the abilities it possesses (or that can be reasonably presumed) to carry out that task. At levels 0 and 1, tasks are defined by promoting fitness and other more immediate goals, such as feeding, mating efforts, avoiding predators, etc. Meanwhile, at level 2, there exists a multitude of tasks, possibly infinite, given that human cognitive capacity allows us to contemplate abstract subjects like mathematical theorems, which may not necessarily be relevant to tasks related to maximizing fitness. Consequently, by focusing solely on level 2 rationality, we encounter an indeterminate number of tasks that could serve as a framework for evaluating rational action and thinking.

At this point, one could object that what has been discussed so far applies primarily to what we might consider as motivational reasons, or at most, reasons based on subjectively given ends. It may not encompass considerations that go beyond individual-level authority, as categorical reasons are supposed to do. In

72 This interactive feedback perspective aligns with our contemporary understanding of the hierarchy of brain areas. For instance, evolutionarily more primitive areas underlying subcortical regions play a role in basic motivation and quick, automatic emotional responses. These areas provide inputs to the cortical regions above them, particularly the prefrontal lobes, which evolved more recently and underlie higher-order cognition. The cortical regions, in turn, respond to impulses and regulate lower-brain areas, forming a loop between higher and lower-level brain regions (see, e.g. Ardila 2008).

response to this objection, the next section will explore considerations that will allow us to broaden the scope of the analysis and accommodate the phenomenology of categorical reasons.

6.4 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: $\exists p \exists q$ RR (If someone believes that p, then she has an intrinsic desire that q)
> R_5: $\exists p$ RR (Rational agents do not desire that p)
> R_6: $\exists q$ RR (Every rational agent desires that q) (Smith 2009, 119–20)

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.[73] 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 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 based on 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. Another example involves the widely accepted claim that if individuals are harmed or injured, then others, if in a position to help, have reasons to assist them.

Unfortunately, the validity of the presented principles is controversial (Smith 2009, 124). Some authors argue for minimal principles of rationality, resembling R_1, while others advocate for more substantive views, allowing principles as strong as R_6. My sense is that the controversy arises, in part, from the belief held by notable authors that if these reason requirements are valid, they need to be justified by *a priori* considerations.

73 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. (see Bermúdez 2003, 110–11; Smith 2012, 234).

For instance, Smith (2012, 238–39) 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 (see 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. Callebaut 2007, 80). However, from a naturalistic point of view, even the *a priori* validity of the instrumental requirement could be challenged.

While the possibility of desiring to achieve a goal might seem conceptually linked to being disposed to take the means believed to be necessary for its accomplishment, it is crucial to distinguish this conceptual connection from the proposed principle of rationality, R_1. According to Smith, for a principle to be considered a principle of rationality, it must guide us on "how to reason when we deliberate" (2009, 121). If R_1 or its variants are norms that one should adhere to in reasoning about what to do, it is conceivable that there are situations in which reasoning in accordance with R_1 may not lead to the fulfillment of one's goals or tasks.

To illustrate this, consider an example given by Jennifer Morton (2011, 569; see, also Broome 2007, 173–74). Imagine a world in which there is an evil demon whose aim is to make your life difficult. Every time you deliberate on the necessary and sufficient means to achieve your goals, the demon alters circumstances to thwart your beliefs from guiding successful actions. Suppose, however, that you are finely attuned to your environment, allowing your instincts to usually lead to successful outcomes. In this scenario, where your perception of affordances is sharp, acting without deliberate consideration often proves successful. In such a world, adherence to instrumental rationality norms would not be beneficial or justified. Instead, relying on instincts emerges as a more effective strategy.

This example highlights the prima facie difficulty in establishing the *a priori* status of instrumental norms of rationality. Given this challenge even for the foundational norm involving means-end reasoning, skepticism arises concerning the feasibility of providing *a priori* justification for other, more substantive norms of rationality. From a naturalistic standpoint, such skepticism is to be expected. In this perspective, our perceived reasons for action and the validity of these beliefs are contingent on factors like our experiences, learning history, cultural background, and reasoning abilities. Additionally, the concept of a rational person is best understood in the context of human rationality, further specified in relation to the tasks humans have evolved to perform phylogenetically or ontogenetically, and their adapted environmental and cultural niches.

To address the existence of categorical reasons, instead of demonstrating how specific norms attained categorical status for individuals, I will present a model that aims to illustrate how this phenomenon could have generically emerged. This approach avoids reliance on *a priori* intuitions regarding the specific reasons we may possess.

6.5 The emergence of categorical reasons

To show how categorical reasons can be accommodated within a naturalistic framework, I will examine how reason relations are likely initially formed. The conclusion drawn from this discussion aims to illustrate that hypothetical and categorical reasons are not inherently different but exist on a continuum, varying in their dependence on individual preferences, beliefs, values, and so forth.

It is crucial to recognize that even at the level of affordances, things presented as favoring a particular course of action are often not phenomenologically construed as contingent on us or subject-based in a broad sense. For instance, when realizing that my life is in danger, I do not perceive the situation as necessi-

tating a response from me because I see myself as an individual with a standing desire or goal to avoid danger. Instead, the typical perspective is that the situation demands a response from us or counts in favor of avoiding danger. Paradoxically, this primitive normativity diminishes when transitioning to the level of reflective rationality, where dispassionate contemplation may lead to inquiries about whether one should avoid danger, adopt a cautious life approach, or take more risks.

Thus, even at this fundamental level, reasons are not depicted as grounded in our subjective needs. Nonetheless, a question persists: when we reach a more reflective level, the normativity of certain situations appears to hinge on our possession of specific desires and goals, while others seem normative irrespective of our individual aims. I maintain that this differentiation between reasons can be elucidated similarly to how naturalistically inclined scholars expound on the formation of semantic relations more broadly. The fundamental concept here is that the establishment of certain primitive semantic relations aligns homomorphically or even isomorphically with establishing particular reason-relations.

6.6 Primitive semantic content and normative reasons

William Harms (2004), relying on Ruth Millikan's (1989) teleological semantic program, constructs a naturalistic framework to account for the genesis of fundamental semantic attributes of indicative and imperative or normative contents within various semantic units. This approach also elucidates the origins of basic normative intuitions concerning the general functioning of things. I suggest extending this framework to the realm of normative reasons.

Within this framework, the central concept is that of primitive content. Primitive content encapsulates representations with a dual purpose: they serve to *indicate* that things possess certain characteristics, while simultaneously signaling which actions

should be undertaken. The idea that certain representations have primitive contents is similar to representations that Millikan (1995) calls 'Pushmi-Pullyu Representations'. Examples illustrating these kinds of representations are often found in the animal kingdom. For instance, the warning calls of vervet monkeys or the distinctive dances of honeybees serve as paradigms. In the case of vervet monkeys, the warning call both indicates the presence of a predator and directs other monkeys to flee. Likewise, honeybees employ a waggle dance that not only indicates the location of foraging or habitat resources but also instructs other bees on the distance and direction to which they should fly.

It is important to note that the fundamental meaning of biological signals, akin to language, is established through convention. Conventions dictate when it is appropriate to emit a signal and specify the suitable action or response corresponding to that signal in a given situation. Take the term "water" as an example; it refers to the H_2O molecule, and its implications include features like transparency, thirst-quenching properties, and suitability for washing. Harms (2004, 193) refers to these two aspects of representations as extension and intension. Extension pertains to what the representation stands for, such as an object or a potential state of affairs, while intensions encompass what results from the proper use of representations within a representational system, determined by their roles and relationships to other representations.

> 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, 194)

As mentioned earlier, representations do not necessarily need to take a linguistic form. In this framework, basic signals like warning cries and bee dances possess meaning, encompassing

both extension and intension. Harms suggests that a representation's content is constituted by the conjunction of its extension and intension. In primitive contents, representations serve both indicative and directive functions. In more sophisticated representations, such as beliefs and desires, these two functions can separate. Beliefs typically have an indicative function, with both extension and intension playing this role, while desires primarily serve a directive function.

For our current discussion, it is crucial to highlight the characteristics that draw parallels between representations and their contents and reasons or facts that support a particular stance. First, representations have extensions, typically regarded as truth-conditions. Similarly, reasons have grounds, encompassing facts, states of affairs, or true propositions that establish the grounds for reason-relations. Second, representations possess intensions, denoting what follows from their role in a representational system concerning the conditions that constitute their extension. Correspondingly, reasons are reasons *for* something, be it an action or an attitude. Third, normative reasons appear to serve a dual purpose. They indicate what appears to be the case while also guiding what should be done in response to the situation. Hence, reasons exhibit features akin to primitive semantic content, simultaneously fulfilling indicative and directive roles.[74] For our current purposes, I suggest equating the reason-relation with representations or a specific subset of representations that exhibit the phenomenology of *counting in favor of* (for a similar suggestion, see Harms and Skyrms 2009, 444–46).

74 We can observe another factor that reinforces the analogy. Just as representations can compete for a response, conflicting reasons, contingent on their weight, can compete for a response. For instance, in a Stroop task, participants are confronted with color words displayed in different colors. The objective is to quickly identify the color of the word. When the word "red" appears in green, individuals often exhibit a bias towards stating that the word is red, even though the actual color is green. This inclination is attributed to the competing representations individuals hold for the same situation.

The analogy between reasons and primitive semantic contents allows us to understand how categorical reasons can be grounded in naturalistic elements. Initially, we can explore the emergence of familiar hypothetical reasons, which depend on the goals of an agent. The establishment of basic semantic relations between signals and responses is commonly explained using a game-theoretical model, as first proposed by David Lewis (1969) and subsequently expanded upon, particularly by Brian Skyrms (1996; 2010). A simple model can depict the establishment of meaning conventions or how a signal acquires specific meaning.

We start by examining a cooperative game with two players or agents.[75] In this game, agents can take on two roles: sender (S) or receiver (R). These roles are not fixed, and an agent may switch between them. Agents can perceive two states of the world (W_1 and W_2), send two messages (M_1 and M_2), and respond with two different actions (A1 and A2). Each action is correct for a specific state of affairs (A_1 for W_1 and A_2 for W_2). If a player correctly responds to a message, both players receive a positive payoff (a>0), otherwise, they receive nothing (payoff is 0). The sender perceives the state of affairs, sends a signal to the receiver, and the goal is achieved if the receiver responds appropriately to the situation, resulting in a positive payoff.

In the absence of a preestablished communication system, four sender and corresponding receiver strategies are available for players to execute in the basic case. These are given in **Figure 4**.

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

75 The following exposition and notation is based on Harms (2004, 194–95) and Huttegger (2007).

Figure 4: Senders can use strategies (S_1-S_4), selecting a message (M) based on conditions (W). Receivers respond with actions (A) based on the received message by using one of the strategies (R1-R4). Nash's equilibrium with maximal payoff is reached when senders using strategies S_1 and S_2 are aligned with receivers' responding by using strategies R_1 and R_2 (adapted from Harms 2004, 194–95; Huttegger 2007).

Agents can combine strategies based on their roles, such as using S_1 as a sender and R_1 as a receiver, or S_2 as a sender and R_2 as a receiver. The possible combinations of strategies are numerous, with 16 different options available. In this example, the focus is on two specific combinations: S_1R_1 and S_2R_2 (see also **Figure 5**), since they bring maximal payoff to the agents (Harms 2004, 195–96). Technically speaking, these combinations (S_1R_1 and S_2R_2) form a Nash equilibrium. In a Nash equilibrium, no single agent has a unilateral incentive to deviate from the established strategies. These specific combinations of strategies achieve this equilibrium by creating a one-to-one relation between states of affairs, messages, and actions. Consequently, if both agents coordinate on either S_1R_1 or S_2R_2, they will consistently benefit from their interactions, ensuring that their responses are optimal given the actions of the other agent. This example highlights the conventional nature of meaning in the established signal system. If players agree on S_1R_1, then M_1 would signify that the world is in state W_1, and A_1 should be executed. Conversely, if they opt for S_2R_2, then M_1 would signify W_2, and A_2 should be performed.

More importantly for the present discussion, this example illustrates how reasons could emerge from interactions between agents. Once the meaning convention is established, in this simple case, a reason-relation is also established. For instance, if S_1R_1 establishes a signal system for when to perform actions A_1 and A_2, then being in W_1 provides a reason or counts in favor of performing A_1 (see **Figure 5**).

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 5: This figure illustrates a reformulation of Figure 4 in terms of reasons. Once a Nash equilibrium is achieved between strategies S_1R_1 or S_2R_2, we can posit the establishment of reason relations that can be recognized or at least experienced as such by (rational) agents participating in the game (adapted from Harms 2004, 196).

The model is naturally applied to interactions between different agents. However, there is nothing inherently preventing its application to single agents. In this sense, the model can explain how particular representations in a single system acquire their meaning or how single reason-relations for specific agents are established. For instance, S_1 can be implemented by an agent's perceptual system, and R_1 as a system that produces actions in response to signals from S_1. When the perceptual system produces signal M_1, an agent would perceive this as a reason or something that counts in favor of performing A_1, whether that is an action or another belief (depending on our interpretation of elements of S_1R_2).

Returning to the interpersonal case, we can observe how categorical reasons can emerge from simple associations between efforts to coordinate actions. Once a sufficient portion of the population adopts the strategy S_1R_1, for instance, it becomes rational for every other agent inclined towards cooperation to regard W_1 as a reason to perform A_1. This holds true regardless of the occurrent preferences or beliefs of that agent. From an evolutionary perspective, the cooperative efforts of many generations of agents will produce a system of reason-relations into which new agents will naturally grow. Many of these reason-relations will be experienced as factors that count in favor of producing appropriate

responses without providing explicit or transparent explanations for why this is the case (for which a detailed examination of the history and evolution of the individual or society of agents would be required) (see Queloz 2021). For example, when we see a person in pain, we understand that she has been hurt and that this situation *demands* a response by helping her in some way. However, the explanation of why this particular fact counts in favor of performing this act will differ depending on the normative narrative that different people accept about the origins or groundings of this relation.

As mentioned earlier, categorical reasons will emerge from interactions between agents, similar to the way reasons emerge at the level of a single agent—by establishing associations between states of affairs and responses that bring some benefit in relation to those states of affairs. However, this will occur only if enough other agents behave in similar ways and adhere to similar associations between states of affairs and actions. In this sense, the emergence of interpersonal categorical reasons will be 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 perspective explain the difference between hypothetical and categorical reasons? The suggestion is that when we engage in reflective thinking, reasons derived from personal goals and desires may appear optional and not externally binding. This intuition could arise because personal reasons depend on contingent plans and desires that are often ephemeral or the products of different quirks. Through self-control, we can influence, change, or come to deem them invaluable.

In contrast, when considering social norms, especially those related to well-being (ours and others'), we think of them as non-optional. We cannot easily influence them solely through self-control; instead, we see them as providing a platform that guides our behavior. This non-optionality (or inescapability)

stems from the fact that social norms are viewed as external to our specific motivational sets. However, when adopting an evolutionary perspective or examining the emergence of reason-relations, there appears to be no qualitative difference between hypothetical and categorical reasons. Categorical reasons at the individual level can be seen as hypothetical at the population level, where agents' strategies reach a stable equilibrium. This is not because an alternative state of affairs (e.g., W_2 instead of W_1) could have been a reason for performing A_1, but rather because the nature of the agents and the interactions between them make certain states of affairs categorical reasons for specific actions.

6.7 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, 206) suggests that normative intuitions are outcomes of higher-order cognitive or affective systems, which take violations of the functions of lower-level systems as inputs and produce responses that reinforce the lower-level rules. In the context of established reason-relations and the mechanisms processing them, higher-level systems regulate and reinforce their functions. For instance, a fundamental requirement for successful cooperation is adherence to the norm of reciprocity (Baumard, André, and Sperber 2013). Agents predisposed to cooperate may intuitively feel a basic reason to expect reciprocity when they perform a significant favor, and *vice versa*. When someone attempts to cheat by receiving a favor without reciprocating, intuitions related to fairness signal a norm violation. These intuitions suggest that punishing such behavior is appropriate, whether through warnings, reporting to authorities, etc. In essence, these intuitions reinforce the basic mechanism handling and satisfying the reason-relation. Similarly, if an individual engages in dishonest behavior, their conscience may produce a reinforcing intuition, with the punitive signal directed inward.

Similarly, epistemic intuitions play a role in regulating how we reason and form beliefs (Harms 2004, 206). This becomes evident when faced with two inconsistent beliefs; the intuition that coherence is violated compels us to abandon one of the beliefs. Typically, the less entrenched belief in our knowledge or belief database is discarded. Epistemic norms also serve a crucial social function (Mercier and Sperber 2011). Communication, with its potential benefits, requires an agent to assess the credibility of information. Instead of relying solely on intuitions, an agent must be capable of evaluating arguments presented by others and generating persuasive arguments that can convince others. This involves overtly rational capacities to properly assess and respond to the available evidence.

In this way, level 2 rationality allows us to respond to already established reason-relations in a more flexible manner compared to automated intuitions. For instance, it enhances our ability to protect ourselves from potential cheaters and to enforce fairness rules more effectively. Moreover, the capacity for detachment from our current motivations and representations, inherent in reflective rationality, empowers us to assess the existing reason-relations we adopt. This enables us to consider whether better normative relations could be established in light of other reasons that we endorse. I will conclude this chapter by illustrating this final point.

Mindless evolutionary processes can lead to many different equilibrium points, establishing various reason-relations. Let us consider a modified signaling game where there is a partial conflict of interest between senders and receivers (Zollman, Bergstrom, and Huttegger 2013).[76] In this situation, senders can be of two types, T_1 and T_2, and they can either send a signal or

76 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, and Huttegger 2013).

not send a signal. The receiver has two possible actions, A_1 and A_2, which 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 against. Therefore, 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 6**). In this game, there is a partial conflict of interest. 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 6: Presented here are sender and receiver strategies, reflecting a partial conflict of interest, with senders categorized as T1 or T2. See the main text for a more detailed explanation (adapted from Zollman, Bergstrom, and Huttegger 2013).

To illustrate how the game functions, we can imagine that senders are people who ask for social benefits and that they differ in their social and economic status. Type T_1 represents those who belong to a lower socio-economic group, and T_2 represents those who belong to a higher socio-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 a situation where there are no signaling costs, it seems that even through 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% of the time 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 the 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 are 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 would they 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 equilibrium 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.

6.8 Concluding remarks

The argument of this book has centered on exploring the nature of normative reasons from the perspective of methodological naturalism. By distinguishing between object- and subject-based theories of normative reasons, it has been argued that the naturalistic perspective aligns more coherently with subject-based theories. This alignment has guided the development of a response-dependence account of normative reasons, emphasizing that the recognition of facts as reasons is influenced by the cognitive and affective makeup of rational agents. Furthermore, the book has examined the implications of evolutionary debunking arguments, proposing that a naturalistic understanding of normativity supports a subject-based theory of normative reasons.

In this final chapter, the goal was to further refine a specific type of subject-based theory of reasons. Specifically, it aimed to demonstrate how categorical reasons might emerge and be explained within this framework. Throughout the chapter, various topics were explored, including the relationship between reasons, the faculty of reason, and norms of rationality. From a naturalistic standpoint, I argued that reasons can be explained in terms of the faculty of reason and the principles that govern it. Within this framework, distinctions were made between different types and criteria of rationality and their relation to reasons. By incorporating a model from game theory, I illustrated how categorical reason-relations could arise. Additionally, I indicated how rationality can be portrayed as a reflective ability that allows agents to detach from their immediate motivations and representations, enabling them to respond to reasons or even establish new rea-

son-relations.

As a final remark, I would like to add that advocating for a subject-based theory of normative reasons has been my way of illustrating how our understanding of reasons is deeply intertwined with the cognitive and affective makeup of individuals and their place in the world. In this regard, the naturalistic approach I have presupposed has guided my exploration and provided a meaningful framework for interpreting the emergence and application of normative reasons. I hope that this work contributes to a deeper understanding of normativity and its role in human reasoning and behavior, and that it inspires further inquiry and reflection in this field that has captivated my attention for so long.

Index

References

Aaltola, Elisa. 2014. "Affective Empathy as Core Moral Agency: Psychopathy, Autism and Reason Revisited." *Philosophical Explorations* 17 (1): 76–92. https://doi.org/10.1080/13869795.2013.825004.

Ainslie, George. 2001. *Breakdown of Will*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139164191.

Alvarez, Maria. 2010. *Kinds of Reasons: An Essay in the Philosophy of Action*. Oxford: Oxford University Press.

———. 2017. "Reasons for Action: Justification, Motivation, Explanation." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2017/entries/reasons-just-vs-expl/.

Andrews, Kristin. 2020. *The Animal Mind: An Introduction to the Philosophy of Animal Cognition*. Second edition. New York and Oxford: Routledge.

Andrews, Kristin, Shannon Spaulding, and Evan Westra. 2021. "Introduction to Folk Psychology: Pluralistic Approaches." *Synthese* 199 (1–2): 1685–1700. https://doi.org/10.1007/s11229-020-02837-3.

Ardila, Alfredo. 2008. "On the Evolutionary Origins of Executive Functions." *Brain and Cognition* 68 (1): 92–99. https://doi.org/10.1016/j.bandc.2008.03.003.

Arkonovich, Steven. 2011. "Advisors and Deliberation." *The Journal of Ethics* 15 (4): 405–24. https://doi.org/10.1007/s10892-011-9101-7.

Arpaly, Nomy, and Timothy Schroeder. 2012. "Deliberation and Acting for Reasons." *The Philosophical Review* 121 (2): 209–39. https://doi.org/10.1215/00318108-1539089.

Asarnow, Samuel. 2019. "Internal Reasons and the Boy Who Cried Wolf." *Ethics* 130 (1): 32–58. https://doi.org/10.1086/704342.

Axelrod, Robert M. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Baccarini, Elvio, and Luca Malatesti. 2017. "The Moral Bioenhancement of Psychopaths." *Journal of Medical Ethics* 43 (10): 697–701. https://doi.org/10.1136/medethics-2016-103537.

Bartels, Daniel M., Christopher W. Bauman, Fiery A. Cushman, David A. Pizarro, and Peter McGraw. 2014. "Moral Judgment and Decision Making." In *Blackwell Reader of Judgment and Decision Making*, edited by G. Keren and G. Wu. Malden, MA: Blackwell.

Baumard, Nicolas, Jean-Baptiste André, and Dan Sperber. 2013. "A Mutualistic Approach to Morality: The Evolution of Fairness by Partner Choice." *Behavioral and Brain Sciences* 36 (1): 59–78. https://doi.org/10.1017/S0140525X11002202.

Baumard, Nicolas, and Pascal Boyer. 2013. "Explaining Moral Religions." *Trends in Cognitive Sciences* 17 (6): 272–80. https://doi.org/10.1016/j.tics.2013.04.003.

Baumard, Nicolas, and Coralie Chevallier. 2015. "The Nature and Dynamics of World Religions: A Life-History Approach." *Proceedings of the Royal Society B: Biological Sciences* 282 (1818): 20151593. https://doi.org/10.1098/rspb.2015.1593.

Baumard, Nicolas, Alexandre Hyafil, Ian Morris, and Pascal Boyer. 2015. "Increased Affluence Explains the Emergence of Ascetic Wisdoms and Moralizing Religions." *Current Biology* 25 (1): 10–15. https://doi.org/10.1016/j.cub.2014.10.063.

Bermúdez, José Luis. 2003. *Thinking without Words*. Oxford: Oxford University Press.

———. 2005. *Philosophy of Psychology: A Contemporary Introduction*. London: Routledge.

Binmore, Ken G. 2007. *Game Theory: A Very Short Introduction*. New York: Oxford University Press.

Blackburn, Simon. 1998. *Ruling Passions: A Theory of Practical Reasoning*. Oxford: Oxford University Press.

Blair, R.J.R. 1995. "A Cognitive Developmental Approach to Morality: Investigating the Psychopath." *Cognition* 57 (1): 1–29. https://doi.org/10.1016/0010-0277(95)00676-P.

Braddock, Matthew. 2016. "Evolutionary Debunking: Can Moral Realists Explain the Reliability of Our Moral Judgments?" *Philosophical Psychology* 29 (6): 844–57. https://doi.org/10.1080/09515089.2016.1163667.

Bratman, Michael. 2007. *Structures of Agency: Essays*. Oxford and New York: Oxford University Press.

Brembs, Björn. 2009. "Mushroom Bodies Regulate Habit Formation in Drosophila." *Current Biology* 19 (16): 1351–55. https://doi.org/10.1016/j.cub.2009.06.014.

Brink, David. 1997. "Kantian Rationalism: Inescapability, Authority, and Supremacy." In *Ethics and Practical Reason*, edited by Garrett Cullity and Berys Nigel Gaut, 255–91. Oxford: Oxford University Press.

Broome, John. 1999. "Normative Requirements." *Ratio* 12 (4): 398–419. https://doi.org/10.1111/1467-9329.00101.

———. 2007. "Is Rationality Normative?" *Disputatio* 2 (23): 161–78. https://doi.org/10.2478/disp-2007-0008.

———. 2013. *Rationality through Reasoning*. Wiley-Blackwell.

———. 2021. "Reason Fundamentalism and What Is Wrong with It." In *Normativity, Rationality and Reasoning*, by John Broome, 11–38. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198824848.003.0002.

Brosnan, Kevin. 2011. "Do the Evolutionary Origins of Our Moral Beliefs Undermine Moral Knowledge?" *Biology and Philosophy* 26 (1): 51–64. https://doi.org/10.1007/s10539-010-9235-1.

Byrne, Alex, and David R. Hilbert. 2003. "Color Realism and Color Science." *Behavioral and Brain Sciences* 26 (1): 3–21. https://doi.org/10.1017/S0140525X03000013.

Call, Josep, and Michael Tomasello. 2008. "Does the Chimpanzee Have a Theory of Mind? 30 Years Later." *Trends in Cognitive Sciences* 12 (5): 187–92. https://doi.org/10.1016/j.tics.2008.02.010.

Callebaut, Werner. 2007. "Herbert Simon's Silent Revolution." *Biological Theory* 2 (1): 76–86. https://doi.org/10.1162/biot.2007.2.1.76.

Chapman, Robert. 2023. *Empire of Normality: Neurodiversity and Capitalism*. London: Pluto Press.

Chen, Yongxiang, Liqi Zhu, and Zhe Chen. 2013. "Family Income Affects Children's Altruistic Behavior in the Dictator Game." Edited by Michel Botbol. *PLoS ONE* 8 (11): e80419. https://doi.org/10.1371/journal.pone.0080419.

Clark, Austen. 2000. *A Theory of Sentience*. New York: Oxford University Press.

Clarke-Doane, Justin. 2020. *Morality and Mathematics*. Oxford: Oxford University Press.

Cline, Brendan. 2015. "Nativism and the Evolutionary Debunking of Morality." *Review of Philosophy and Psychology* 6 (2): 231–53. https://doi.org/10.1007/s13164-014-0207-2.

Colyvan, Mark. 2009. "Naturalising Normativity." In *Conceptual Analysis and Philosophical Naturalism*, edited by David Braddon-Mitchell and Robert Nola, 303–13. Cambridge, Mass.: MIT Press.

Copp, David. 1995. *Morality, Normativity, and Society*. Oxford: Oxford University Press.

Cullity, Garrett. 2022. "Reasons and Fit." In *Fittingness*, edited by Chris How-

ard and R. A. Rowland, 151–75. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780192895882.003.0007.

Dancy, Jonathan. 2000. *Practical Reality*. Oxford: Oxford University Press.

———. 2004. *Ethics without Principles*. Oxford: Oxford University PressOxford. https://doi.org/10.1093/0199270023.001.0001.

Davidson, Donald. 2001. *Essays on Actions and Events*. Clarendon: Oxford University Press. https://doi.org/10.1093/0199246270.001.0001.

Dawkins, Richard. 1986. *The Selfish Gene*. Oxford: Oxford University Press.

De Caro, Mario. 2023. "Between the Placement Problem and the Reconciliation Problem. Philosophical Naturalism Today." *Topoi* 42 (3): 675–82. https://doi.org/10.1007/s11245-023-09913-6.

Dennett, Daniel C. 2003. *Freedom Evolves*. London: Penguin Books.

Doris, John M., and Stephen P. Stich. 2011. "As a Matter of Fact: Empirical Perspectives on Ethics." In *Knowledge, Rationality, and Morality, 1978-2010*, by Stephen P. Stich, 114–52. Collected Papers 2. Oxford: Oxford university press.

Dreier, Jamie. 2015. "Can Reasons Fundamentalism Answer the Normative Question?" In *Motivational Internalism*, edited by Gunnar Björnsson, Caj Strandberg, Ragnar Francén Olinder, John Eriksson, and Fredrik Björklund, 167–81. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199367955.003.0009.

Edmonds, David. 2023. *Parfit: A Philosopher and His Mission to Save Morality*. Princeton: Princeton University Press.

El Mouden, Claire, Jean-Baptiste. André, Olivier Morin, and Daniel Nettle. 2014. "Cultural Transmission and the Evolution of Human Behaviour: A General Approach Based on the Price Equation." *Journal of Evolutionary Biology* 27 (2): 231–41. https://doi.org/10.1111/jeb.12296.

El Mouden, Claire, Maxwell Burton-Chellew, Andy Gardner, and Stuart A. West. 2012. "What Do Humans Maximise?" In *Evolution and Rationality: Decisions, Co-Operation and Strategic Behaviour*, edited by Samir Okasha and Ken Binmore, 23–49. Cambridge: Cambridge University Press.

Enoch, David. 2005. "Why Idealize?" *Ethics* 115 (4): 759–87. https://doi.org/10.1086/430490.

———. 2010. "The Epistemological Challenge to Metanormative Realism: How Best to Understand It, and How to Cope with It." *Philosophical Studies* 148 (3): 413–38. https://doi.org/10.1007/s11098-009-9333-6.

———. 2011. *Taking Morality Seriously: A Defense of Robust Realism*. Oxford: Oxford University Press.

Field, Hartry. 2009. "What Is the Normative Role of Logic?" *Aristotelian Society Supplementary Volume* 83 (1): 251–68. https://doi.org/10.1111/j.1467-8349.2009.00181.x.

Fischer, John Martin, and Mark Ravizza. 2000. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Fitzpatrick, William. 2008. "Robust Ethical Realism, Non-Naturalism, and Normativity." In *Oxford Studies in Metaethics*, edited by Russ Shafer-Landau, 3:159–205. Oxford: Oxford University Press.

FitzPatrick, William J. 2016. "Misidentifying the Evolutionary Debunkers' Error: Reply to Mogensen." *Analysis* 76 (4): 433–37. https://doi.org/10.1093/analys/anw065.

Fogal, Daniel, and Olle Risberg. 2023. "Explaining Normative Reasons." *Noûs* 57 (1): 51–80. https://doi.org/10.1111/nous.12393.

Foot, Philippa. 1972. "Morality as a System of Hypothetical Imperatives." *Philosophical Review* 81 (3): 305–16.

Frankfurt, Harry G. 1988a. "Identification and Wholeheartedness." In *The Importance of What We Care about: Philosophical Essays*, by Harry G. Frankfurt, 159–76. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511818172.

———. 1988b. "The Importance of What We Care About." In *The Importance of What We Care about: Philosophical Essays*, 80–94. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511818172.

Garson, Justin. 2015. *The Biological Mind: A Philosophical Introduction*. Abingdon, Oxon: Routledge.

Gaus, Gerald F. 2011. *The Order of Public Reason*. Cambridge: Cambridge University Press.

Gettier, Edmund. 1963. "Is Knowledge Justified True Belief?" *Analysis* 23 (6): 121–23.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Oxford and New York: Oxford University Press.

Gibson, James J. 1979. *The Ecological Approach to Visual Perception: Classic Edition*. Boston: Houghton Mifflin.

Giere, Ronald N. 2006. *Scientific Perspectivism*. Chicago: University of Chicago Press.

———. 2008. "Naturalism." In *The Routledge Companion to Philosophy of Science*, edited by Stathis Psillos and Martin Curd, 213–23. London and New York: Routledge.

Goldman, Alan H. 2007. "Desire, Depression, and Rationality." *Philosophical Psychology* 20 (6): 711–30. https://doi.org/10.1080/09515080701665912.

———. 2009. *Reasons from within: Desires and Values*. Oxford: Oxford University Press.

Goodwin, Geoffrey P., and John M. Darley. 2008. "The Psychology of Meta-Ethics: Exploring Objectivism." *Cognition* 106 (3): 1339–66. https://doi.org/10.1016/j.cognition.2007.06.007.

Haidt, Jonathan. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108 (4): 814–34.

———. 2007. "The New Synthesis in Moral Psychology." *Science* 316 (5827): 998–1002. https://doi.org/10.1126/science.1137651.

Hardin, C. L. 1988. *Color for Philosophers: Unweaving the Rainbow*. Indianapolis: Hackett Pub. Co.

———. 2003. "A Spectral Reflectance Doth Not a Color Make:" *Journal of Philosophy* 100 (4): 191–202. https://doi.org/10.5840/jphil2003100420.

Hare, Robert D. 1993. *Without Conscience: The Disturbing World of the Psychopaths among Us*. New York: Guilford Press.

Harman, Gilbert H. 1977. *The Nature of Morality: An Introduction to Ethics*. New York: Oxford University Press.

———. 2000. "Is There a Single True Morality?" In *Explaining Value and Other Essays in Moral Philosophy*. Oxford: Oxford University Press.

———. 2004. "Practical Aspects of Theoretical Reasoning." In *The Oxford Handbook of Rationality*, edited by Alfred R. Mele and Piers Rawling, 45–56. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195145397.003.0003.

Harms, William F. 2004. *Information and Meaning in Evolutionary Processes*. New York: Cambridge University Press.

Harms, William F., and Brian Skyrms. 2009. "Evolution of Moral Norms." In *The Oxford Handbook of Philosophy of Biology*, edited by Michael Ruse, 434–50. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195182057.003.0019.

Hippel, William von, and Robert Trivers. 2011. "The Evolution and Psychology

of Self-Deception." *Behavioral and Brain Sciences* 34 (1): 1–16. https://doi.org/10.1017/S0140525X10001354.

Hirstein, William, Katrina L. Sifferd, and Tyler Fagan. 2018. *Responsible Brains: Neuroscience, Law, and Human Culpability*. Cambridge, Massachusetts: The MIT Press.

Hopster, Jeroen. 2018. "Evolutionary Arguments against Moral Realism: Why the Empirical Details Matter (and Which Ones Do)." *Biology & Philosophy* 33 (5): 41. https://doi.org/10.1007/s10539-018-9652-0.

———. 2019. "Striking Coincidences: How Realists Should Reason about Them." *Ratio* 32 (4): 260–74. https://doi.org/10.1111/rati.12221.

———. 2020. "Explaining Historical Moral Convergence: The Empirical Case against Realist Intuitionism." *Philosophical Studies* 177 (5): 1255–73. https://doi.org/10.1007/s11098-019-01251-x.

Huemer, Michael. 2016. "A Liberal Realist Answer to Debunking Skeptics: The Empirical Case for Realism." *Philosophical Studies* 173 (7): 1983–2010. https://doi.org/10.1007/s11098-015-0588-9.

Huttegger, Simon M. 2007. "Evolutionary Explanations of Indicatives and Imperatives." *Erkenntnis* 66 (3): 409–36.

Hutto, Daniel D., and Ian Ravenscroft. 2021. "Folk Psychology as a Theory." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2021/entries/folkpsych-theory/.

Jackson, Frank. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Clarendon.

Jefferson, Anneli, and Katrina L. Sifferd. 2018. "Are Psychopaths Legally Insane?" *European Journal of Analytic Philosophy* 14 (1): 79–96. https://doi.org/10.31820/ejap.14.1.5.

Johnston, Mark. 1989. "Dispositional Theories of Value." *Proceedings of the Aristotelian Society, Supplementary Volumes* 63:139–74.

———. 1992. "How to Speak of the Colors." *Philosophical Studies* 68 (3): 221–63. https://doi.org/10.1007/BF00694847.

Joyce, Richard. 2006. *The Evolution of Morality*. Cambridge, Mass.: The MIT Press.

———. 2013. "The Evolutionary Debunking of Morality." In *Reason and Responsibility*, edited by Joel Feinberg and Russ Shafer-Landau, 527–34. Boston: Wadsworth Cengage Learning.

Jurjako, Marko. 2017. "Agency and Reductionism about the Self." In *Perspec-*

tives on the Self, edited by Boran Berčić, 255–84. Rijeka: University of Rijeka.

———. 2022. "Naturalizam i relativnost u pogledu praktičnih razloga (Engl. Naturalism and Relativism about Practical Reasons)." In *Human Rationality: Festschrift for Nenad Smokrović*, edited by Boran Berčić, Aleksandra Golubović, and Majda Trobok, 113–39. Rijeka: University of Rijeka, Faculty of Humanities and Social Sciences.

Jurjako, Marko, and Luca Malatesti. 2016. "Instrumental Rationality in Psychopathy: Implications from Learning Tasks." *Philosophical Psychology* 29 (5): 717–31.

Kacelnik, Alex. 2006. "Meanings of Rationality." In *Rational Animals?*, edited by Susan Hurley and Matthew Nudds, 87–106. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198528272.003.0002.

Kahane, Guy. 2011. "Evolutionary Debunking Arguments." *Noûs* 45 (1): 103–25. https://doi.org/10.1111/j.1468-0068.2010.00770.x.

Kennett, Jeanette, and Cordelia Fine. 2009. "Will the Real Moral Judgment Please Stand Up?" *Ethical Theory and Moral Practice* 12 (1): 77–96. https://doi.org/10.1007/s10677-008-9136-4.

Kiesewetter, Benjamin. 2017. *The Normativity of Rationality*. Oxford: Oxford University Press.

Kitcher, Philip. 2011. *The Ethical Project*. Cambridge, Mass.: Harvard University Press.

Knauff, Markus, and Wolfgang Spohn. 2021. *The Handbook of Rationality*. Cambridge, Mass.: The MIT press.

Kolodny, Niko. 2005. "Why Be Rational?" *Mind* 114 (455): 509–63. https://doi.org/10.1093/mind/fzi509.

Korsgaard, Christine M. 1986. "Skepticism about Practical Reason." *Journal of Philosophy* 83 (1): 5–25.

———. 1996. *The Sources of Normativity*. Cambridge; New York: Cambridge University Press.

———. 2011. "The Activity of Reason." In *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon*, edited by R. Jay Wallace, Rahul Kumar, and Samuel Freeman, 3–22. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199753673.001.0001.

Krebs, Dennis. 2005. "The Evolution of Morality." In *The Handbook of Evolutionary Psychology*, edited by D. Buss. Hoboken, NJ: Wiley-Blackwell.

———. 2011. *The Origins of Morality: An Evolutionary Account*. New York: Oxford University Press.

Kumar, Victor, and Richmond Campbell. 2022. *A Better Ape: The Evolution of the Moral Mind and How It Made Us Human*. New York, NY: Oxford University Press.

Laland, Kevin N. 2008. "Exploring Gene-Culture Interactions: Insights from Handedness, Sexual Selection and Niche-Construction Case Studies." *Philosophical Transactions of the Royal Society B: Biological Sciences* 363 (1509): 3577–89. https://doi.org/10.1098/rstb.2008.0132.

Lawson, David W., and Monique Borgerhoff Mulder. 2016. "The Offspring Quantity–Quality Trade-off and Human Fertility Variation." *Philosophical Transactions of the Royal Society B: Biological Sciences* 371 (1692): 20150145. https://doi.org/10.1098/rstb.2015.0145.

Lekić Barunčić, Kristina. 2021. *Philosophical Perspectives on Autism*. Rijeka: Faculty of Humanities and Social Sciences.

Lenman, James. 2005. "The Saucer of Mud, the Kudzu Vine and the Uxorious Cheetah: Against Neo-Aristotelian Naturalism in Metaethics." *European Journal of Analytic Philosophy* 1 (2): 37–50.

———. 2009. "Reasons for Action: Justification vs. Explanation." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2009. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/Archives/spr2009/entries/reasons-just-vs-expl/.

Levy, Arnon, and Yair Levy. 2018. "Evolutionary Debunking Arguments Meet Evolutionary Science." *Philosophy and Phenomenological Research*. https://doi.org/10.1111/phpr.12554.

Lewis, David K. 1969. *Convention: A Philosophical Study*. Cambridge, Mass.: Harvard.

Lillehammer, H. 2000. "Revisionary Dispositionalism and Practical Reason." *The Journal of Ethics* 4 (3): 173–90. https://doi.org/10.1023/A:1009822006329.

Logins, Artūrs. 2022. *Normative Reasons: Between Reasoning and Explanation*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009076012.

López De Sa, Dan. 2013b. "Rigid vs Flexible Response-Dependent Properties." In *Varieties of Dependence: Ontological Dependence, Grounding, Supervenience, Response-Dependence*, edited by Miguel Hoeltje, Benjamin Schnieder, and Alex Steinberg, 393–418. München: Philosophia Verlag. https://doi.org/10.2307/j.ctv2nrzhj9.

———. 2013a. "The Aposteriori Response-Dependence of the Colors." *Croatian Journal of Philosophy* 13 (1): 65–79.

Lord, Errol. 2018. *The Importance of Being Rational*. Oxford: Oxford University Press.

Machuca, Diego E., ed. 2023. *Evolutionary Debunking Arguments: Ethics, Philosophy of Religion, Philosophy of Mathematics, Metaphysics, and Epistemology*. New York, NY: Routledge.

Mackie, John Leslie. 1977. *Ethics: Inventing Right and Wrong*. New York: Penguin Books.

Maibom, Heidi L. 2018. "What Can Philosophers Learn from Psychopathy?" *European Journal of Analytic Philosophy* 14 (1): 63–78.

Manne, Kate. 2014. "Internalism about Reasons: Sad but True?" *Philosophical Studies* 167 (1): 89–117. https://doi.org/10.1007/s11098-013-0234-3.

Mantel, Susanne. 2018. *Determined by Reasons: A Competence Account of Acting for a Normative Reason*. New York: Routledge. https://doi.org/10.4324/9781351186353.

Markovits, Julia. 2014. *Moral Reason*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199567171.001.0001.

May, Joshua. 2014. "Does Disgust Influence Moral Judgment?" *Australasian Journal of Philosophy* 92 (1): 125–41. https://doi.org/10.1080/00048402.2013.797476.

Mayr, Ernst. 1961. "Cause and Effect in Biology." *Science* 134 (3489): 1501–6.

Međedović, Janko. 2023. *Evolutionary Behavioral Ecology and Psychopathy*. Cham: Springer. https://doi.org/10.1007/978-3-031-32886-2.

Mercier, Hugo, and Dan Sperber. 2011. "Why Do Humans Reason? Arguments for an Argumentative Theory." *Behavioral and Brain Sciences* 34 (2): 57–74. https://doi.org/10.1017/S0140525X10000968.

Millgram, Elijah. 1995. "Was Hume a Humean?" *Hume Studies* 21 (1): 75–94.

Millikan, Ruth Garrett. 1989. "Biosemantics." *The Journal of Philosophy* 86 (6): 281–97. https://doi.org/10.2307/2027123.

———. 1995. "Pushmi-Pullyu Representations." *Philosophical Perspectives* 9:185–200. https://doi.org/10.2307/2214217.

Miščević, Nenad. 2004. "Response-Intentionalism about Color: A Sketch." *Croatian Journal of Philosophy* 4 (2): 179–91. https://doi.org/10.5840/croatjphil20044211.

———. 2007. “Is Color-Dispositionalism Nasty and Unecological?” *Erkenntnis* 66 (1): 203–31. https://doi.org/10.1007/s10670-006-9036-8.

Mogensen, Andreas L. 2015. “Evolutionary Debunking Arguments and the Proximate/Ultimate Distinction.” *Analysis* 75 (2): 196–203. https://doi.org/10.1093/analys/anv013.

———. 2016. “Do Evolutionary Debunking Arguments Rest on a Mistake about Evolutionary Explanations?” *Philosophical Studies* 173 (7): 1799–1817. https://doi.org/10.1007/s11098-015-0579-x.

Mollon, J. D. 1989. ““Tho’ She Kneel’d in That Place Where They Grew…” The Uses and Origins of Primate Colour Vision.” *Journal of Experimental Biology* 146 (1): 21–38. https://doi.org/10.1242/jeb.146.1.21.

Morton, Jennifer M. 2011. “Toward an Ecological Theory of the Norms of Practical Deliberation.” *European Journal of Philosophy* 19 (4): 561–84. https://doi.org/10.1111/j.1468-0378.2010.00400.x.

Nettle, Daniel, Agathe Colléony, and Maria Cockerill. 2011. “Variation in Cooperative Behaviour within a Single City.” Edited by Yamir Moreno. *PLoS ONE* 6 (10): e26922. https://doi.org/10.1371/journal.pone.0026922.

Nichols, Shaun. 2002. “On the Genealogy of Norms: A Case for the Role of Emotion in Cultural Evolution.” *Philosophy of Science* 69 (2): 234–55. https://doi.org/10.1086/341051.

Nowak, Martin A. 2006. “Five Rules for the Evolution of Cooperation.” *Science* 314 (5805): 1560–63. https://doi.org/10.1126/science.1133755.

Nuccetelli, Susana, and Gary Seay, eds. 2012. *Ethical Naturalism: Current Debates*. Cambridge: Cambridge University Press.

Olson, Jonas. 2014. *Moral Error Theory: History, Critique, Defence*. Oxford: Oxford University Press.

O’Neill, Onora. 1996. “Introduction.” In *The Sources of Normativity*, by Christine M Korsgaard, xi–xv. Oxford: Oxford University Press.

Palmer, Stephen E. 1999. *Vision Science: Photons to Phenomenology*. Cambridge, Mass.: MIT Press.

Papineau, David. 2002. *Thinking about Consciousness*. Oxford: Oxford University Press. https://doi.org/10.1093/0199243824.001.0001.

———. 2023. “Naturalism.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Fall 2023. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2023/entries/naturalism/.

Parfit, Derek. 1997. "Reasons and Motivation." *Aristotelian Society Supplementary Volume* 71 (1): 99–130. https://doi.org/10.1111/1467-8349.00021.

———. 2011a. *On What Matters*. Vol. 1. Oxford and New York: Oxford University Press.

———. 2011b. *On What Matters*. Vol. 2. Oxford and New York: Oxford University Press.

———. 2017. *On What Matters*. Vol. 3. Oxford: Oxford University Press.

Peacocke, Christopher. 1992. *A Study of Concepts*. Cambridge, Mass.: MIT Press.

Pollock, John L. 1987. "Defeasible Reasoning." *Cognitive Science* 11 (4): 481–518. https://doi.org/10.1207/s15516709cog1104_4.

Pölzler, Thomas. 2018. "How to Measure Moral Realism." *Review of Philosophy and Psychology* 9 (3): 647–70. https://doi.org/10.1007/s13164-018-0401-8.

———. 2023. *A Philosophical Perspective on Folk Moral Objectivism*. New York: Routledge.

Prinz, Jesse. 2006. "The Emotional Basis of Moral Judgments." *Philosophical Explorations* 9 (1): 29–43.

Pukka, Bill. 2023. "The Golden Rule." In *The Internet Encyclopedia of Philosophy*. https://www.iep.utm.edu/goldrule/.

Queloz, Matthieu. 2021. *The Practical Origins of Ideas: Genealogy as Conceptual Reverse-Engineering*. Oxford: Oxford University Press.

Quine, Willard van Orman, ed. 1981. *Theories and Things*. Cambridge, Mass.: Harvard University Press.

Railton, Peter. 1986. "Moral Realism." *The Philosophical Review* 95 (2): 163. https://doi.org/10.2307/2185589.

———. 2004. "How to Engage Reason: The Problem of Regress." In *Reason and Value: Themes From the Moral Philosophy of Joseph Raz*, edited by R. Jay Wallace, Philip Pettit, Samuel Scheffler, and Michael Smith. Oxford: Clarendon Press.

———. 2009. "Practical Competence and Fluent Agency." In *Reasons for Action*, edited by David Sobel and Steven Wall, 81–115. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511720185.005.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.

Raz, Joseph. 1975. *Practical Reason and Norms*. Oxford: Oxford University

Press.

———. 1999. *Engaging Reason: On the Theory of Value and Action*. New York: Oxford University Press.

Reisner, Andrew Evan, and Asbjørn Steglich-Petersen, eds. 2011. *Reasons for Belief*. Cambridge: Cambridge University Press.

Richerson, Peter J., and Robert Boyd. 2005. *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago: University of Chicago Press.

Roberts, Deborah. 2005. "Does the Explanatory Constraint on Practical Reasons Favour Naturalism about Practical Reasons?" *South African Journal of Philosophy* 24 (2): 97–108. https://doi.org/10.4314/sajpem.v24i2.31417.

Rosenberg, Alexander. 2011. *The Atheist's Guide to Reality: Enjoying Life without Illusions*. New York, NY: Norton.

Ross, W. D. 1930. *The Right and the Good*. Clarendon Press.

Rowland, Richard. 2019. *The Normative and the Evaluative: The Buck-Passing Account of Value*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198833611.001.0001.

Ruse, Michael. 1986. "Evolutionary Ethics: A Phoenix Arisen." *Zygon* 21 (1): 95–112. https://doi.org/10.1111/j.1467-9744.1986.tb00736.x.

Ruse, Michael, and Edward O. Wilson. 1986. "Moral Philosophy as Applied Science." *Philosophy* 61 (236): 173–92. https://doi.org/10.1017/s0031819100021057.

Sackris, David, and Rasmus Rosenberg Larsen. 2023. "Are There 'Moral' Judgments?" *European Journal of Analytic Philosophy* 19 (2): (S1)1-23. https://doi.org/10.31820/ejap.19.2.1.

Samuels, Richard, Stephen Stich, and Michael Bishop. 2002. *Ending the Rationality Wars: How to Make Disputes about Human Rationality Disappear*. Edited by Renee Elio. *Common Sense, Reasoning, and Rationality*. Oxford: Oxford University Press.

Sarkissian, Hagop, John Park, David Tien, Jennifer Cole Wright, and Joshua Knobe. 2011. "Folk Moral Relativism." *Mind & Language* 26 (4): 482–505. https://doi.org/10.1111/j.1468-0017.2011.01428.x.

Scanlon, Thomas M. 1998. *What We Owe to Each Other*. Cambridge, Mass.: Harvard University Press.

———. 2014. *Being Realistic about Reasons*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199678488.001.0001.

Schafer, Karl. 2015a. "Hume on Practical Reason: Against the Normative Au-

thority of Reason." In *Oxford Handbook of David Hume*, edited by Paul Russell. Oxford: Oxford University Press.

———. 2015b. "Realism and Constructivism in Kantian Metaethics (1): Realism and Constructivism in a Kantian Context." *Philosophy Compass* 10 (10): 690–701. https://doi.org/10.1111/phc3.12253.

———. 2015c. "Realism and Constructivism in Kantian Metaethics (2): The Kantian Conception of Rationality and Rationalist Constructivism." *Philosophy Compass* 10 (10): 702–13. https://doi.org/10.1111/phc3.12252.

———. 2018. "A Brief History of Rationality: Reason, Reasonableness, Rationality, and Reasons." *Manuscrito* 41 (4): 501–29. https://doi.org/10.1590/0100-6045.2018.v41n4.ks.

Schroeder, Mark A. 2007. *Slaves of the Passions*. Oxford: Oxford University Press.

———. 2008. "Having Reasons." *Philosophical Studies* 139 (1): 57–71.

———. 2021. *Reasons First*. Oxford: Oxford University Press.

Searle, John R. 2001. *Rationality in Action*. Cambridge, Mass.: MIT Press.

Severini, Eleonora. 2016. "Evolutionary Debunking Arguments and the Moral Niche." *Philosophia* 44 (3): 865–75. https://doi.org/10.1007/s11406-016-9708-9.

Shafer-Landau, Russ. 1999. "Moral Judgement and Normative Reasons." *Analysis* 59 (1): 33–40. https://doi.org/10.1093/analys/59.1.33.

———. 2003. *Moral Realism: A Defence*. Oxford and New York: Oxford University Press.

———. 2012. "Evolutionary Debunking, Moral Realism, and Moral Knowledge." *Journal of Ethics and Social Philosophy* 7 (1): 1–37.

Simon, Herbert A. 1956. "Rational Choice and the Structure of the Environment." *Psychological Review* 63 (2): 129–38. https://doi.org/10.1037/h0042769.

Singer, Peter. 2017. *Does Anything Really Matter? Essays on Parfit on Objectivity*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199653836.001.0001.

Sinnott-Armstrong, Walter P., ed. 2008. *Moral Psychology (Vols. i-iii)*. Cambridge, Mass.: MIT Press.

Skorupski, John. 2010. *The Domain of Reasons*. Oxford: Oxford University Press.

Skyrms, Brian. 1996. *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

———. 2010. *Signals: Evolution, Learning, and Information*. Oxford: Oxford University Press.

Smith, Michael. 1987. "The Humean Theory of Motivation." *Mind* 96 (381): 36–61.

———. 1994. *The Moral Problem*. Oxford: Blackwell.

———. 2004. *Ethics and the a Priori: Selected Essays on Moral Psychology and Meta-Ethics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511606977.

———. 2009. "Desires, Values, Reasons, and the Dualism of Practical Reason." *Ratio* 22 (1): 98–125. https://doi.org/10.1111/j.1467-9329.2008.00420.x.

———. 2012. "Naturalism, Absolutism, Relativism." In *Ethical Naturalism: Current Debates*, edited by Susana Nuccetelli and Gary Seay, 226–44. Cambridge; New York: Cambridge University Press.

———. 2013. "A Constitutivist Theory of Reasons: Its Promise and Parts." *Law, Ethics and Philosophy*, no. 1, 9–30.

Smokrović, Nenad. 2018. "Informal Reasoning and Formal Logic: Normativity of Natural Language Reasoning." *Croatian Journal of Philosophy* 18 (54): 455–70.

Sobel, David. 2009. "Subjectivism and Idealization." *Ethics* 119 (2): 336–52. https://doi.org/10.1086/596459.

Sober, Elliott. 2000. *Philosophy of Biology*. 2. ed. Boulder, Colorado: Westview Press.

Sober, Elliott, and David Sloan Wilson. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. 4. print. Cambridge, Mass.: Harvard University Press.

Starzak, Tobias, and Tobias Schlicht. 2023. "Can Affordances Be Reasons?" *Philosophical Psychology*, 1–27. https://doi.org/10.1080/09515089.2023.2270694.

Stearns, Stephen C. 2004. *The Evolution of Life Histories*. Oxford: Oxford University Press.

Stich, Stephen P. 1990. *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation*. Cambridge, Mass.: MIT Press.

Street, Sharon. 2006. "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127 (1): 109–66.

———. 2008a. "Constructivism about Reasons." *Oxford Studies in Metaethics* 3:207–45.

———. 2008b. "Reply to Copp: Naturalism, Normativity, and the Varieties of Realism Worth Worrying About." *Philosophical Issues* 18 (1): 207–28. https://doi.org/10.1111/j.1533-6077.2008.00145.x.

———. 2009. "In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters." *Philosophical Issues* 19:273–98.

———. 2017. "Nothing 'Really' Matters, but That's Not What Matters." In *Does Anything Really Matter*, edited by Peter Singer, 121–48. Oxford: Oxford University Press.

Streumer, Bart. 2023. *Unbelievable Errors: An Error Theory about All Normative Judgements*. Oxford: Oxford University Press.

Sun, Yifan. 2022. "Response-Dependence and Normativity." *Theoria* 88 (6): 1128–43. https://doi.org/10.1111/theo.12429.

Sušnik, Matej. 2015. "Strong Motivational Internalism." *International Philosophical Quarterly* 55 (2): 165–77. https://doi.org/10.5840/ipq201542031.

Sylvan, Kurt. 2015. "What Apparent Reasons Appear to Be." *Philosophical Studies* 172 (3): 587–606.

Taccolini, Joshua. 2024. "Can We Defend Normative Error Theory?" *European Journal of Analytic Philosophy* 20 (1): 131–54. https://doi.org/10.31820/ejap.20.1.6.

Trivers, Robert L. 1971. "The Evolution of Reciprocal Altruism." *The Quarterly Review of Biology* 46 (1): 35–57. https://doi.org/10.1086/406755.

Turner, Stephen P. 2010. *Explaining the Normative*. Cambridge: Polity.

Tye, Michael. 2002. *Consciousness, Color, and Content*. Cambridge, Mass.: MIT.

Ventham, Elizabeth. 2023. "Morality without Categoricity." *European Journal of Analytic Philosophy* 19 (2): (A4)1-23. https://doi.org/10.31820/ejap.19.2.4.

Verbeek, Bruno. 2007. "The Authority of Norms." *American Philosophical Quarterly* 44 (3): 245–58.

———. 2008. "Conventions and Moral Norms: The Legacy of Lewis." *Topoi* 27 (1–2): 73–86. https://doi.org/10.1007/s11245-008-9029-0.

Way, Jonathan. 2017. "Reasons as Premises of Good Reasoning." *Pacific Philosophical Quarterly* 98 (2): 251–70. https://doi.org/10.1111/papq.12135.

Wedekind, Claus, and Manfred Milinski. 2000. "Cooperation through Image Scoring in Humans." *Science* 288 (5467): 850–52.

Wedgwood, Ralph. 2007b. "Normativism Defended." In *Contemporary Debates in Philosophy of Mind*, edited by Brian P. McLaughlin and Jonathan D. Cohen, 85–102. Oxford: Blackwell.

———. 2007a. *The Nature of Normativity*. Oxford and New York: Clarendon Press; Oxford University Press.

Williams, Bernard. 1981. "Internal and External Reasons." In his *Moral Luck*, 101–13. Cambridge: Cambridge University Press.

———. 1995. "Internal Reasons and the Obscurity of Blame." In his *Making Sense of Humanity*, 35–45. Cambridge: Cambridge University Press.

Wilson, David Sloan, Daniel Tumminelli O'Brien, and Artura Sesma. 2009. "Human Prosociality from an Evolutionary Perspective: Variation and Correlations at a City-Wide Scale." *Evolution and Human Behavior* 30 (3): 190–200. https://doi.org/10.1016/j.evolhumbehav.2008.12.002.

Wimmer, Heinz, and Josef Perner. 1983. "Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception." *Cognition* 13 (1): 103–28. https://doi.org/10.1016/0010-0277(83)90004-5.

Wright, Crispin. 1992. *Truth and Objectivity*. Cambridge Mass.: Harvard University Press.

Wright, Jennifer C., Piper T. Grandjean, and Cullen B. McWhite. 2013. "The Meta-Ethical Grounding of Our Moral Beliefs: Evidence for Meta-Ethical Pluralism." *Philosophical Psychology* 26 (3): 336–61. https://doi.org/10.1080/09515089.2011.633751.

Wright, Wayne. 2013. "Color Constancy Reconsidered." *Acta Analytica* 28 (4): 435–55. https://doi.org/10.1007/s12136-013-0187-3.

Zollman, Kevin J. S., Carl T. Bergstrom, and Simon M. Huttegger. 2013. "Between Cheap and Costly Signals: The Evolution of Partially Honest Communication." *Proceedings of the Royal Society B: Biological Sciences* 280 (1750): 20121878. https://doi.org/10.1098/rspb.2012.1878.

In "Normative Reasons from a Naturalistic Point of View", Marko Jurjako explores the foundational concept of normative reasons through the lens of methodological naturalism. Departing from traditional analytic or purely conceptual approaches, this philosophical inquiry navigates the terrain of reasons, rationality, and normativity—concepts with a long philosophical pedigree—within a framework rooted in our understanding of the natural and social world. By aligning the exploration with scientifically based theorizing, the book seeks a synoptic view that bridges the gap between relatively isolated philosophical discussions of the nature of normative reasons and their grounding in human intrapersonal and interpersonal interactions that can be understood and explained by an array of scientifically inspired models.

Starting off from Derek Parfit's framework distinguishing between object-based and subject-based theories, Jurjako takes readers on a thought-provoking journey through the intricacies of the ontology of normative reasons. Drawing upon insights from game theory, cognitive sciences, and evolutionary theory, the book weaves a narrative that defends a view according to which normative reasons are fundamentally based on rational individuals' practical natures and their interactions with other agents.

While primarily designed for philosophers and graduate students working at the intersection of normativity and cognitive sciences, the book should be accessible to curious readers from diverse fields eager to grasp the nature of normative reasons and their connection to human rational capacities. As such, "Normative Reasons from a Naturalistic Point of View" invites you to explore a fresh perspective on the nature of reasons.

Marko Jurjako is an Associate Professor in the Department of Philosophy and the Division of Cognitive Sciences at the Faculty of Humanities and Social Sciences, University of Rijeka. He is the Editor-in-Chief of the *European Journal of Analytic Philosophy* and currently serves as the President of the Croatian Society for Analytic Philosophy.

9 789533 611327

ONLINE BOOK VERSION