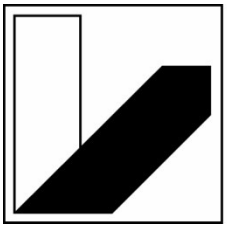

UNIVERSITÄT BAYREUTH

**FACULTY OF CULTURAL STUDIES**

—

**DEPARTMENT OF PHILOSOPHY**

# ORIGINS OF MORAL RELEVANCE

The Psychology of Moral Judgment, and its Normative and Metaethical Significance

A dissertation submitted
in partial fulfillment of the requirements
for the degree of
Doktor der Philosophie (Dr. phil.)

Written under the direction of
Prof. Dr. Rainer Hegselmann

Prepared by

Benjamin Huppert M.A.

Reviewers

Prof. Dr. Rainer Hegselmann
Prof. Dr. Matthew Braham

Submitted 10 June, 2015

Accepted 15 July, 2015

# Acknowledgements

This dissertation was submitted to the Faculty of Cultural Studies at the University of Bayreuth on June 10, 2015 and accepted on July 15, 2015. I am deeply grateful for the friendship, love, and kindness I received while I was working on it. My family has been an indispensable source of strength, and my supervisor Rainer Hegselmann provided support far beyond the ordinary. My colleagues at Bayreuth's Department of Philosophy created an atmosphere of curiosity, good humor, and encouragement. Thank you Eckhart Arnold, Matthew Braham, Alexander Brink, Uwe Czaniera, Huimin Dong, Annette Dufner, Ben Ferguson, Claudia Ficht, Julian Fink, Carola Friedel, Roberto Fumagalli, Brigitte Goßler, Niels Gottschalk-Mazouz, Marcel Kiel, Carlo Martini, Olivier Roy, Rudolf Schüßler, Attila Tanyi, Jan-Willem van der Rijt, and Sonja Weber. Friends elsewhere gave me confidence by taking an interest in my progress. Thank you Ani Andree, Cora Beck, Max Blüher, Sacha Droste, Nermana Dzino, Mathis Eisenhardt, Valerie Ficke, Sabine Fuß, Michael Goldhammer, Jenny Grimm, Susanne Harms, Michael Haun, Katrin & Nina Heinzmann, Andreas Jost, Steffi Jünger, Serap Kilicaslan, Katharina Labermeier, Martin Lunge, Silvia Maier, Pascal Partikel, Marianna Spatola, Emanuele Terracini, and Sophie Wiesner. Some individuals have been particularly involved with my toiling one way or another. Lea Diederichsen, Lars Eickmeier, Gordian Haas, Stephanie Hartmann, Anna Ovcharenko, Jan Reiter, Oliver Will, and Joachim Wündisch: Without you, this would have been a great deal harder. Finally, I owe thanks to the students of Bayreuth's Philosophy and Economics program for reinforcing my fascination with the science of morality, to Joshua Greene for taking the time to answer some of my questions, and to the participants of numerous workshops and conferences for providing input and feedback. I am standing on the shoulders of all of you.

Düsseldorf, June 2017

# Contents

# List of Figures



# List of Tables

Significant parts of this dissertation report findings from social and cultural psychology, cognitive science, neuroscience, and primatology. Where is the philosophy in that? I consider attention to empirical research indispensable to addressing the questions with which I am concerned, lest my philosophical views be incompatible with how science sees morality. In my view, such incompatibility greatly diminishes the relevance of moral philosophy.

# Introduction

The science of morality is on the rise. Cognitive and neurological science, evolutionary psychology, social and cultural psychology, evolutionary biology, and other disciplines are generating more and more findings and theories about the onto- and phylogenesis of morality, as well as about the details of moral cognition. The emerging picture indicates that moral cognition, judgment, and action involve many different psychological processes that interact in complex patterns. Rather detached from these endeavors, moral concepts like right and wrong, questions regarding how one should conduct one's life, justice, fairness, etc., are part of our daily experience. Even though such notions may be implicit and vague, most people harbor some normative ideas about which actions morality requires, descriptive ideas about the origin of these requirements or recommendations (e.g., divine commandments, social convention, etc.), and more or less elaborate notions of moral concepts. The origin and content of morality are also essential topics of moral philosophy: Is morality based on reason or emotion? Is it universal or relative to culture or even individuals? Can we know what is right and wrong, and if so, how?

A scientific understanding of morality relates to both philosophical and folk notions of morality. Both contain descriptive and normative propositions. Science's bearing on descriptive notions of morality is relatively straightforward. Its relevance for normative notions, however, is a more complicated subject. Scientific findings and accounts of morality can relate to each other in various ways: For instance, scientific findings could be incompatible with descriptive parts of a specific notion of morality. Normative moral philosophical theories might prove too demanding given actual human capacity, or they might mistake a subset of phenomena within the realm of morality for the whole of morality. Information about the development of different moralities could prompt people to reevaluate their moral beliefs or become more tolerant of other views. Because the science of morality appears to deal with concepts that also figure in philosophical or folk notions, several findings

have been taken to 'make a difference' to simple as well as to more sophisticated conceptions of morality.

One particular class of supposedly significant findings identifies 'unexpected' features of moral cognition and morally relevant action.[1] Moral judgment and behavior, it seems, are surprisingly emotion-driven, intuitive, susceptible to the influence of situational features, shaped by the way in which the human brain evolved, and possibly heuristic. These results point to numerous processes involved in moral judgment, and the positions that regard empirical results as philosophically significant typically aim to assess the adequacy of judgments that share one or several of these characteristics. The findings are frequently taken to identify conditions under which ordinary moral judgment is unreliable, or even generally inadequate. However, evaluating the adequacy of moral judgment and action requires a standard by which adequacy can be judged. In my view, a conception of how moral judgment *should* work, if at all conceivable, needs to be based on an understanding of what kind of phenomena morality and moral judgment actually are. I believe that setting standards for moral judgment and action without comprehending the psychology underlying them is a misguided project. Many relevant psychological processes are part of human nature and thus not easily deactivated. Moreover, standard setting *for* moral judgments might itself be executed by the mechanisms that are also *involved in* moral judgment to a substantive degree. If the appropriateness of judgments is questioned based on information about the genesis and functioning of the processes generating these judgments, and if the standards by which we assess appropriateness involve similar processes, it is necessary to question also the standards of appropriateness themselves. In order to find out whether the same or similar processes are indeed involved in both moral judgments and evaluations of the adequacy of moral judgments, we need a psychological account of both.

Since reactions to research on morality can be viewed as psychological phenomena, I believe that conjectures regarding the impact of findings about morality have to be based on moral psychology, just like the evaluation of different kinds of moral judgments. I argue that the picture of morality and moral judgment shaped by recent efforts in its scientific investigation *is* significant for moral philosophy. In my opinion, moral psychology renders mind-independent accounts of morality rather implausible. Moreover, the various features of moral judgments that supposedly undermine their reliability are in fact among the origins of our most fundamental moral beliefs. These conclusions emerge from the discussion of

[1] To avoid this cumbersome expression, I will use 'moral action' to refer to deeds with a moral dimension, i.e., moral or immoral acts.

inadequacy accusations leveled against moral judgments based on their determinants, and of other positions that hold that the corresponding findings have no or only indirect normative significance. In order to arrive at these results, I will provide a sketch of moral psychology that is corroborated by illustrative findings from several other disciplines.

The global structure of the dissertation is as follows: Part I opens with a short primer on twentieth-century moral psychology, an account of the rise of evolutionary thought in psychology, and a presentation of some philosophically provocative findings from contemporary moral psychology. The second element of Part I is a detailed presentation and preliminary analysis of selected philosophical responses to these and related results. I argue that proper philosophical assessment requires a deeper psychological understanding of morality in general and the concept of moral relevance in particular, since such relevance is a crucial element in many of the proposed arguments. Providing this understanding is the aim of Part II. More specifically, I discuss psychological theories of the moral domain, the complex relations between emotions and morality, the mechanics and heritability of moral cognition, and the causes of moral disagreement that these perspectives suggest. Part III reconsiders the philosophical suppositions of Part I in the light of the descriptive account of morality developed in Part II, resulting in my own conjectures about the repercussions of a psychological understanding of morality in moral philosophy.

# Part I
—
# Traditional Moral Psychology and Philosophical Reactions to New Findings

# 1 From Traditional to Evolutionary Moral Psychology, and Provocative Findings[2]

## 1.1 Moral Psychology up to the Mid-Twentieth Century

Even before psychology became a discipline in its own right, theories of society, morality, and moral judgment contained psychological elements. For instance, such theories made claims about what *motivates* human beings to judge and act morally (self-interest vs. sympathy; merely aversive or also positive emotions, etc.), the nature of the *mental processes* which determine moral evaluations (emotional vs. rational; conscious vs. unconscious), or about whether moral norms are products of culture rather than human biology. My aim in this chapter is not to engage in exegesis of these positions, but merely to set the stage for the ensuing, more fine-grained discussions of contemporary moral-psychological findings.

The philosopher Thomas Hobbes held that distinctions between good and evil only come into existence once individuals transfer some of their powers onto a sovereign out of self-interest. On his account, sympathy, or a general interest in fellow human beings, is not a required motive for the establishment of a moral order. Crucially, Hobbes posited that human nature contains few traits that are conducive to peaceful coexistence. Rather, man has to overcome the shortsightedness of the self-interested motives ingrained in his nature using reason and is fit for social living only by conscious decisions to act *against* his natural tendencies. The view that morality consists in *overcoming* the behavioral dispositions with which human beings are naturally equipped resonates in the writings of biologist Thomas Huxley, who was otherwise an ardent vindicator of Charles Darwin's theory of evolution.[3] To Huxley, morality was inexplicable by reference to evolutionary processes, which he thought promote only narrowly self-interested behavioral tendencies. Frans de Waal, a prominent primatologist, coined the term *veneer theory of morality* for a tradition of positions that "sees people as essentially evil and selfish and explains morality as a cultural overlay ungrounded in human nature or evolutionary theory."[4] De Waal notes a connection between this tradition and early psychological accounts of morality:

---

[2] This dissertation marks the preliminary completion of a philosophical project begun with my Master's thesis. The following chapters have (distant) ancestors or contain thoughts present in Huppert (2010): 1.1.2, 1.3, 1.3.1, 1.3.3, 2.1, 2.8, 3, 3.1, 3.2, 3.2.1, 4.1, 4.3.1, 4.3.5, 4.4.1.2, 4.4.1.3, 4.4.2, 4.4.3, 4.5, 5, 5.1, 5.2, 5.3, 5.4, 5.5, 6, 6.1, 9.2, 10.

[3] See De Waal (2005), p. 18. Huxley did not, however, have formal education and "did not accept natural selection as the chief engine of evolution." De Waal (2013), p. 34.

[4] De Waal (2005), p. 31.

> Huxley's dualism was to get a respectability boost from Sigmund Freud's writings, which thrived on contrasts between the conscious and subconscious, the ego and super-ego, Love and Death, and so on. […] [Freud] let civilization arise out of a renunciation of instinct, the gaining of control over the forces of nature, and the building of a cultural super-ego.[5]

Regarding the nature of moral judgment, there is a related opposition between two schools of thought. *Rationalism* holds that reason is decisive in the *detection* of adherence to or violation of norms, in the gauging and classification of such behaviors, and the formulation of judgment. Immanuel Kant is considered to be the epitome of this approach, Plato a predecessor, and John Rawls its most prominent modern proponent. Cognitive-developmental models of moral psychology as devised by Jean Piaget and Lawrence Kohlberg, explained below, also form part of this tradition: Their understanding of competence in moral judgment is closely tied to confidence in the use of conscious deliberation.[6]

According to the so-called *sentimentalist* view of morality, on the other hand, moral judgment is essentially affect laden, and humans are frequently seen as motivated not only by self-interest, but also by a genuine concern for the well-being of others.[7] Scottish philosophers David Hume and Adam Smith are regarded as its most formative advocates. Sentimentalists believe that moral judgment originates in emotional responses which are not necessarily affected by conscious reasoning and tend to arise automatically (intuitively); reasoning operates on the categories of moral relevance delineated by affective responses. Darwin and Freud can also be counted among the sentimentalists.[8] Jonathan Haidt, whose theories on moral judgment and the domain of morality play a major role in this dissertation, is one of the preeminent contemporary sentimentalist moral psychologists.

### *1.1.1 Freud, Behaviorism, and Social Learning Theory on Morality*

Within the psychoanalytical tradition, as in much of twentieth-century psychology, morality as such was not a central issue. In line with the pivotal role of therapy in psychoanalysis, its main concerns were pernicious effects of an overdeveloped super-ego, supposedly the locus of moral norms internalized from the parents, on the individual, and socially detrimental effects of an imperfect super-ego.[9] Freud saw morality as the regulator of conflict between

5 Ibid., p. 18.
6 See Greene (2002), p. 218.
7 Jesse Prinz's term, used in a talk in Barchem/Netherlands in August 2011.
8 See Damasio (2005), p. 53 for this account of the rationalist/sentimentalist divide.
9 See Sunar (2009), pp. 449–450.

individual desires and the requirements of societal existence that results from "the incompatibility of psychological and biological needs of individuals and strivings for long-term survival of individuals and the species."[10] Within this framework, moral development proceeds via the child's acquisition of society's moral norms fueled by the psychological "dynamics of the Oedipal conflict."[11] The struggle between individual desires and the requirements of social living is mirrored in the conflict between the Id and the super-ego. Freud's theory posits several important characteristics of morality that influenced later psychological thought: Morality is a necessary condition of social existence to which the individual must adhere, even if its demands run counter to her desires. Freud accorded a central role to emotions. Not only did he hold that emotions like guilt keep behavior in check, but he also thought that moralization is a strategy for dealing with the unpleasant experience of anxiety and jealousy. Unlike Hume, Freud assumed that individuals are fundamentally separate: Other individuals matter because of one's own needs, not out of genuine concern with their well-being.[12]

Behaviorism focused on processes of learning and memory more generally. Thus, its account of morality is mostly an application of principles of learning to a particular subject matter. As in Freudian theory, aversive emotions like fear, shame, guilt and anxiety figure prominently in behaviorist models of the acquisition of moral norms, though the mechanisms posited are different: Behaviorism holds that the mind is devoid of any content at birth and is 'written upon' through the mechanisms of classical and operative conditioning. According to B. F. Skinner, moral behavior is essentially the kind of behavior that is reinforced by value judgments based on societal norms.[13] This position has been taken to imply that actions are neither good nor bad 'in themselves'.[14] Social Learning Theory shared with behaviorism the notion of the mind as a blank slate, and the assumption of general learning mechanisms. Within this framework, however, norms are internalized mainly through the imitation of behavioral examples, as well as in response to punishment and reward. Conditioned anxiety is the only emotion of major importance in this process.[15]

10 Turiel (2006a), p. 790.
11 Sunar (2009), p. 448.
12 See ibid. for this characterization of Freud's take on morality.
13 See Lapsley (1996), p. 42.
14 See Turiel (2006a), p. 790 and ibid., pp. 799–800.
15 See Sunar (2009), p. 448.

### *1.1.2 The Cognitive-Developmental Tradition: Piaget and Kohlberg*

In the previous section, I classified Piaget and Kohlberg's prototypical cognitive-developmental accounts of moral acquisition as congenial to a rationalist conception of morality. Piaget and Kohlberg investigated the psychological processes underlying morality by focusing on its development in individuals.[16] Both conceptualized morality as essentially concerned with rules that regulate social interaction; moral judgment evaluates adherence to or deviation from such rules. They were concerned mainly with *reasoning* about moral issues and held that individual moral development consists in the advancement of such reasoning capabilities in an invariable sequence.

Piaget posited three successive levels in the understanding of rules: rules as individual rituals, heteronomous morality, and autonomous morality. Each stage contains elements of the preceding stages, but integrates them in a more stable and more extensive manner.[17] According to Piaget, children acquire a mature understanding of moral concepts not so much from the moral guidelines provided by parents or other authority figures (heteronomous morality/*morality of constraint*), since they achieve obedience mainly by relying on their superior physical, intellectual, and social power. Rather, children develop a more autonomous understanding of morality by interacting with peers:

> It is through cooperative exchanges with agemates that children come to learn that rules are not blind requests for obedience but are instead socially constructed flexible arrangements that serve pragmatic ends and are binding as long as consensus prevails, mutual interests are served, and the bonds of solidarity protected. Because social power is more evenly distributed within a peer group, peers must be won over with reasons.[18]

Piaget focused on the development of conscious moral deliberation. Rather than being conditioned to obey norms through emotions elicited by reward and punishment, Piaget depicted the child as constructing a mature morality via active engagement with its environment (its peers, in particular) and conscious analysis of these interactions in terms of reciprocity, intentionality, equality, social institutions, rules, and authority.[19] He had "little concern for either the emotional aspects emphasized by Freud or the role of direct inculcation of morality envisioned by social learning theory."[20] While actions are compelled by acquired

16 See Heidbrink (2008), pp. 57–87 for an extensive account of Piaget's and Kohlberg's theories.
17 See Kohlberg (1973), p. 632.
18 Lapsley (1996), p. 17.
19 See Sunar (2009), p. 449 and Turiel (2006a), pp. 790–791.
20 Sunar (2009), p. 449.

habit in behaviorism and prescribed by an internalized super-ego in Freud's theory, Piaget's notion of autonomy allows for a more active role of the individual. Moreover, moral evaluations are taken to be sensitive to social context and somewhat flexible.[21]

Kohlberg's related, more fine-grained framework distinguishes three levels of development, each of which contains two stages characterized by specific kinds of justification given for the assessment of various moral dilemmas. Children at the first (preconventional) level justify rule compliance by reference to punishment and private interests: "You ought not to do A, otherwise you will be punished." On the second, 'conventional' level, children aim to comply with whatever rules are emphasized by attachment figures and attempt to uphold the social order: "You ought not to do A, because X, Y and Z do not approve of it/because it is against the law." Only the subgroup of individuals who ascends to the two last (postconventional) stages reasons from general principles (justice, in particular) and criticizes extant rules on that basis.

> […] the three levels can be thought of as three ways of relating the self to the moral expectations of society. In the preconventional level, moral rules and norms are external to persons; that is, they are imposed from the outside by authority figures. At the conventional level, the self internalizes the expectations of authority. At the postconventional level, what is outside (expectations of authority and of society) and what is inside (self-chosen principles) are clearly distinguished, with emphasis placed on the latter for defining moral options.[22]

Kohlberg, like Piaget, rejects behaviorist and social-learning-theoretical models of morality as something introduced into the individual from the outside. His conception of a postconventional level of moral deliberation, however, took the momentum even further: It emphasized the possibility of critical reflection on extant norms enabled by taking the perspective of others. In Kohlberg's view, this capacity implies that there is a universal morality, which can be cognized through reflection and experience.[23] The critical abilities attained at the postconventional level also include the capacity to distinguish alterable, *conventional* rules, from more fundamental *moral* rules.[24] The distinction between moral and conventional norms will serve as a first psychological delineation of the moral domain in chapter 3.1.

---

21 See Turiel (2006a), p. 791.
22 Lapsley (1996), p. 67.
23 See Sunar (2009), p. 449.
24 This does not mean that *only* individuals on the highest Kohlbergian stages *distinguish* between moral and conventional rules. However, only they derive this distinction from general principles and use these principles to determine which rules to apply in a given situation.

For much of the twentieth century, Kohlberg's perception of the moral domain as concerned mainly with issues of justice, rights, and the harm individuals inflict on each other as well as his focus on conscious, verbal moral reasoning dominated moral psychology. In the meantime, however, an alternative perspective was slowly growing from roots in evolutionary biology, until, in the last decades of the twentieth century, it began to investigate phenomena with which the cognitive-developmental moral psychologists had been concerned.

## 1.2 Evolutionary Psychology: Origins and Concepts

> Throughout most of the 20th century, the psychology of morality was dominated by three theories – psychoanalytic theory, social learning theory, and cognitive developmental theory – but in recent years, it has been transformed by a veritable explosion of new concepts and theories, touched off in large part by the emergence of evolutionary psychology.[25]

This chapter explains core concepts of evolutionary thought and subsequently introduces fundamental notions in evolutionary psychology. Section 1.2.1 sketches central ideas in evolutionary theory and psychology. Section 1.2.2 outlines the development of evolutionary psychology; 1.2.3 presents the theory of modular brain architecture, which holds that the mind contains many separately evolved psychological mechanisms. It also addresses some worries about evolutionary psychology. In its entirety, this chapter sets forth the theoretical background underlying the notions of moral cognition discussed subsequently.

### *1.2.1 Central Tenets of Evolutionary Theory*

The central tenet of evolutionary theory is that species change, and it suggests principles that govern this process. Frequently, evolution is defined as change of the relative frequencies of genes within a population over the course of generations.[26] Charles Darwin's *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life* (1859) introduced two basic ideas: the 'tree of life', and natural selection. The 'tree of life' illustrates the relatedness of all species; they descend from a common ancestor like branches from a tree's stem. Natural selection is one of several mechanisms that explain why the tree grows different branches. However, while other mechanisms (genetic drift[27]) also cause evolutionary change, only natural selection accounts for the fact that creatures appear to be

[25] Ibid., p. 447.
[26] See Sober (2000), p. 1.
[27] Changes in the frequency of an allele due to random effects.

*designed* for survival and reproduction in specific environments.[28] Darwin observed that no two living beings are exactly alike, and that the differences between them can have effects on the likelihood of reproduction. Natural selection occurs if variation between individuals leads to changes in relative reproductive success: An antelope that runs faster than its conspecifics survives longer (on average) because it can outrun predators more often, and thus, other things being equal, enjoys more opportunities to mate. Characteristics like 'greater running speed' increase their possessor's (classical[29]) *fitness*, i.e., the likelihood of individual reproductive success, or the relative frequency of its genes in the next generation.[30] If these traits are hereditary, evolutionary change by natural selection ensues: The antelope passes on her running abilities; her offspring equally reaps the corresponding increase in fitness. Over several generations advantageous traits spread, disadvantageous traits are crowded out. The accumulation of these effects spawns substantive change in a species' functional design and, in time, new species.[31]

Design features that became permanent because they increased fitness are called *adaptations*. Adaptations solve specific *adaptive problems*[32] reliably and are one of three evolutionary categories of traits. Adaptive problems are tasks that had to be solved repeatedly in order to survive and reproduce, such as finding food, avoiding predators, or securing a mate. The features of an individual's environment that constitute an adaptive problem make up the corresponding adaptation's *environment of evolutionary adaptedness* (EEA).[33] Evolutionary psychology assumes that many behavioral traits observable today were formed in an EEA that differed from present day environments.[34] It is important to note that many adaptations take quite some time to evolve; they develop only in response to somewhat permanent features of the environment.[35] Thus, the anatomical design of modern humans is thought to be adapted primarily to selective pressures prevalent during the Pleistocene (roughly 1.8 million to 11500 years ago), a time in which our ancestors lived a nomadic life. Humans are considered to have been 'anatomically modern' for 200000 years. Cultural modernity, in

28 See Pinker (2001), p. 468.
29 Cf. 'inclusive fitness' below.
30 Variation need not be hereditary for natural selection to occur. Heritability is, however, necessary for *evolution* by natural selection.
31 See Futuyma (2005), pp. 6–7, Sober (2000), pp. 18–22.
32 See Voland (2007), p. 24.
33 See Buss (2008), p. 40.
34 See Mascaro et al. (2010), p. 15.
35 Note that while it remains undisputed that many adaptations took tens of thousands of years to form, recent findings suggest that adaptations can also emerge much more rapidly. Well-known examples include the development of lactose tolerance in humans and changes in the metabolism of starch; the claim also gains credibility from the significant changes made to wildlife species in the process of their domestication.

contrast, as indicated by the presence of variable and complex artifacts, was achieved only 45000 years ago.[36] If ecological circumstances change as rapidly as some of the conditions of human existence have during the last few hundred years, these adaptations can remain advantageous, but they can also become neutral or even disadvantageous. Consequently, adaptations are not necessarily *adaptive* (conducive to inclusive fitness[37]) at all times, in every environment, and not every trait that is adaptive in specific circumstances is an adaptation. Saying that something is an *adaptation* is a claim about the development of that trait and a claim about its long-term fitness effects in the EEA; saying that a trait is *adaptive* is a primarily a statement about the fitness effects of that trait in specific circumstances. These two properties can come apart.[38] The distinction between adaptations and adaptiveness enlightens the understanding of a second class of traits: *by-products*. In contrast to adaptations, these traits did not spread because they solved adaptive problems in the EEA. Rather, they are regular side effects of adaptations that were either neutral (on average) with respect to fitness or not detrimental enough to offset the benefits of the corresponding adaptation during the period in which that adaptation developed. Just like adaptations, by-products can be regularly adaptive or maladaptive in environments that differ from the EEA or in specific circumstances. The navel of placental mammals, for instance, is a by-product of the umbilical-cord solution (adaptation) to the problem of feeding the embryo or fetus in the womb. Both adaptations and by-products are species typical (all normal members of a species have them). A third category of traits, *noise*, comprises *individual* traits that are neither adaptations nor by-products, such as the unique shape of a navel.[39]

### *1.2.2 Towards Evolutionary Psychology*

Darwin's theory explained adaptation through natural selection, but provided no satisfactory account of heredity. Only by the 1930s-1940s, the so-called *new synthesis* wedded evolutionary theory to the principles of genetics discovered by Gregor Mendel. The 1960s and 1970s brought another significant advance: Previously, evolutionary processes were analyzed in terms of the advantages specific traits bestow upon *individuals* or *groups of individuals*. From around 1964 onwards, several biologists (William Hamilton, George C. Williams, and

36 See Boehm (2012), p. 82.

37 Classical fitness refers to the reproductive success of an individual, while inclusive fitness takes a gene's-eye view and refers to the reproductive success of specific alleles that are present in individuals and their relatives.

38 See James (2011), pp. 21–23.

39 See Buss (2008), pp. 39–42. Such individual traits can nevertheless be (mal)adaptive or neutral in specific environments.

Richard Dawkins, among others) reinterpreted Darwin's theory from a gene's-eye point of view and thereby broadened the explanatory scope of natural selection. From this perspective, it is not just the effect on its carrier's *individual* fitness that determines whether a heritable trait spreads. Rather, successful traits increase their relative frequency in a population by generating more copies of the *genes* that encode them than competing designs do. Copies of genes, not individuals, persist through replication. Hamilton's concept of *inclusive fitness*[40] captures this rationale: In addition to the classical fitness of the trait-carrying individual, it encompasses the fitness of its genetic relatives, since relatives *share* genes with certain probabilities depending on degree of kinship. The corresponding process is called *kin selection*, and it provides an explanation for a phenomenon that classical fitness alone could not explain. Dispositions for altruistic behavior, i.e., behavior to the disadvantage of the executing organism but to the benefit of another, can prevail if they increase the relative frequency of the genes causing them, for instance if a sacrifice saves relatives that share the gene(s) in question with a probability sufficient to offset the loss.[41]

The case of altruism illustrates an important point: Natural selection explains not only the development of physical traits, but accounts also for some species-typical patterns of *behavior*. *Sociobiology*, established as a distinct research program in the 1970s mainly through the work of Edward O. Wilson, applied evolutionary thinking specifically to the study of *social* behavior, and pioneered the interpretation of *human* behavior along these lines.[42] It shared this focus on behavior with *behaviorism,* the dominant psychological paradigm in the United States during the time of sociobiology's inception. In time, the sociobiological method of explaining human behavioral tendencies by reference to principles of natural selection spurred an interest in morality within psychology, where it had so far been a rather peripheral subject. Moral behavior came to intrigue researchers because, at first glance, it is at odds with traits like selfishness or aggression whose contribution to survival and procreation is evident.[43]

Morality was not, however, of central concern to behaviorists. They analyzed psychological phenomena in terms of observable input (stimulus) from the environment and observable behavioral output (response). Since the processes generating output from input were unobservable, conjectures about them (the 'black box' of the mind) were dismissed as

---

40 See ibid., pp. 13–14.

41 See for instance Cosmides & Tooby (1992), pp. 167–169. According to a so-called slippage model, this process might also account for a limited degree of extra-familial helping behavior due to imprecision of kin recognition, as long as the costs are not too high. See Boehm (2012), p. 57.

42 See Kitcher (1985), pp. 113–116.

43 See Sunar (2009), pp. 449–450.

unscientific. For behaviorists, the human mind was a 'blank slate', equipped by nature merely with a general propensity to *learn* anything by conditioning, including all abilities required for survival.[44] Behaviorism prevailed between roughly the 1920s and the 1970s. While psychologists influenced by evolutionary theory such as Sigmund Freud or William James had devised theories of the human psyche in which *innate instincts* figured prominently, behaviorism turned away from that notion since instincts were unobservable internal states and because, to behaviorists, complex behavior was not inherited but the result of learning.[45] When the new behaviorists (Skinner) encountered difficulty explaining certain behaviors without referring to instincts, they introduced the concept of 'drives'. They suggested that humans and animals are homeostatic mechanisms, and that drives are signals that initiate action to maintain homeostasis.[46] In the late 1960s and early 1970s, however, findings emerged which could not be reconciled with the postulates of behaviorism. It seemed that nonhuman animals and humans were 'hardwired' to learn some things more easily than others, and exhibited behavior that was inexplicable by operant conditioning alone. In a famous experiment conducted in 1966, John Garcia and Robert Koelling irradiated rats that had just fed in order to nauseate them. After a single trial, the rats had learned to avoid the kind of food that had been paired with the radiation. However, when the sickening radiation co-occurred with light flashes or buzzes, the rats did not learn to avoid these stimuli.[47] Other studies found that it is much easier to instill into humans a fear of snakes and spiders than of power sockets or cars, even though the latter cause far more deaths in modern environments.[48] These unexpected results indicated that the mind might not be quite as blank a slate as behaviorists had thought: Contrary to the behaviorist thesis of equipotentiality, not just any two stimuli can equally be associated through learning. Consequently, scientific interest in the *processes* leading from stimulus to response increased. This shift in focus, dubbed the 'cognitive revolution' of psychology, was inspired by concepts from computer technology. The human mind came to be regarded as an information processor, programmed to handle specific types of information in specific ways. The shift from behaviorism to cognitive approaches also affected the view of the human mind in another, related aspect: In the heyday of behaviorism, scientific psychologists rejected 'mentalism' and avoided writing about both conscious and unconscious mental processes. Even when they

44 See Buss (2008), p. 28 or Pinker (2006), p. 2.
45 See Schacter et al. (2009), p. 389.
46 See ibid., p. 390.
47 See Buss (2008), p. 30.
48 See ibid. Some authors claim that aversion felt towards these animals is based on disgust rather than fear. See Rozin et al. (2008), p. 760.

became more concerned with what was going on in people's minds, they hesitated to employ terms like 'consciousness' or 'unconsciousness' for fear of being considered Freudians juggling fuzzy concepts, as opposed to proper scientists.[49] Instead, they wrote about processes that were 'implicit', 'pre-attentive', 'procedural', or 'automatic'.[50]

Today, the unconscious once again figures in the writings of eminent psychologists, particularly those with an evolutionary mindset. The modern conception of the 'adaptive unconscious' is, however, quite different from the Freudian unconscious. While for Freud what was unconscious was not conscious because it was *repressed*, contemporary approaches posit that many important functions of the mind operate unconsciously because their emergence dates back to times when consciousness did presumably not yet exist, and because it was evolutionarily *advantageous* for humans to perform many kinds of information processing rapidly and automatically. Our conscious capacities are insufficient to accomplish all the processing necessary for survival (proprioception, color vision etc.).[51] Many automatic, unconscious processes are more ancient than conscious processes. Similar or equivalent processes are discernible in both contemporary animal relatives and our evolutionary ancestors, while the emergence of consciousness appears to be a relatively recent phenomenon typically attributed exclusively to humans. *Dual-process models of cognition*, which distinguish automatic, unconscious functioning from controlled, conscious processes, are at the center of some important debates in contemporary moral psychology that I will discuss.

Unlike recent approaches in evolutionary psychology, early cognitive psychology retained the notion of general-purpose cognition prevalent in behaviorism. While behaviorism held that humans possess a general *learning instinct*, early cognitive psychology postulated a general *information processing mechanism*. From an evolutionary point of view, the idea of such domain-general mechanisms is questionable: Typically, we understand the physiology of evolved species as an assembly of separate, yet tightly integrated organs that fulfill different functions. Lungs and livers, for instance, fulfill disparate functions. The popularity of this modular understanding of the body is certainly aided by the fact that, contrary to details of neural circuitry, many organic structures outside the skull are discernible to the naked eye.[52] Evolutionary theory and the concept of natural selection provide a rationale for this functional anatomic architecture: Organs developed because they helped solve our ancestors' adaptive problems. These problems were highly diverse; thus, organs are diverse. Livers

49 See Wilson (2002), p. 4.
50 See ibid., p. 5.
51 This is true for both lower- and higher-level processes. See ibid., p. 8.
52 See Hagen (2005), pp. 154–155.

process harmful substances, eyes serve orientation in the environment, etc. If this organizational pattern of specialization and integration characterizes the rest of the body, why should the brain be different? Thinking of the evolution of the brain as analogous to the evolution of other organs leads to the observation that the problems solved by mental processing are as diverse as those with which other organs deal are. Since there does not seem to be a 'general adaptive problem' handled by the mind, why assume a general mental mechanism? Evolutionary psychology takes mental activity and capacities to correspond to physical phenomena and structures in the brain and is thus opposed to 'Cartesian dualism', the rather strict division of body and mind associated with René Descartes.[53] From this perspective, in some sense, the body (brain) *is* the mind. Accordingly, if the mind is modular, then the brain is probably organized in a modular fashion as well and vice versa. This notion does allow that multiple brain structures are involved in the operation of a single mental module, or that individual brain structures play a role in various mental modules. Importantly, the view that neural phenomena constitute the material substrate of mental processes does not imply that all mental phenomena are genetically determined. The environment exerts significant influence on the occurrence of specific mental phenomena via long-term effects on brain development and via current 'input' to neurological processes. Nevertheless, from an evolutionary perspective, it is likely that many features of the human mind are adaptations.

The fact that neural structures are very small might have hindered a 'modular' understanding of the brain in the past. However, technology is making progress, and even so, thinking about adaptive problems in our predecessors' environment is a useful strategy to formulate hypotheses about psychological phenomena since "[n]atural selection has mapped the structure of the environment onto the structure of organisms."[54] Rather than assume a general information processing mechanism, evolutionary psychology posits the existence of multiple specialized *evolved psychological mechanisms* (EPMs) or modules, that is, psychological adaptations.[55]

### *1.2.3 Evolved Psychological Mechanisms*

Evolutionary psychology rests on the assumption that the brain is the physical substrate of the mind, and that evolutionary processes shaped the brain just as they shaped the rest of human physiology. It brings to bear on the brain the "scientific model of the body as set of

53 See Wilson (2002), p. 9.
54 Hagen (2005), p. 155.
55 See Buss (2008), p. 50 and Mascaro et al. (2010), p. 15.

[…] distinct mechanisms that function to enable and facilitate the survival and reproduction of the individual organism."[56] While evolutionary psychology has inherited from sociobiology the idea that natural selection shapes behavior, the notion of psychological mechanisms adds an important new element: "Human behaviors are not a direct product of natural selection but rather the product of psychological mechanisms that were selected for."[57]

As argued in the preceding section, understanding natural selection as competition in solving various adaptive problems leads to the expectation that the mind contains functionally separate mechanisms, and that separate neural phenomena correspond to these mechanisms. For instance, processes responding to the sight of a snake presumably differ from those involved in evaluating potential mates. Evidence from functional brain imaging and examinations of individuals suffering from localized brain damage support this *modularity hypothesis*: Specific tasks generate specific patterns of activity in certain parts of the brain; patients with locally damaged neural tissue loose particular abilities that are retained in cases with other lesions.[58] Apparently, distinct areas of the brain serve different functions, as do separate organs in the rest of the body. Although we are still far from understanding exactly how neuronal networks represent and process information, it seems there are many, potentially evolved psychological mechanisms engraved in specific neural patterns that qualify as adaptations.[59] Note that the modularity hypothesis does not hold that *each* evolved module corresponds to a *single* behavioral trait. Like polygenic traits (traits that involve several genes), certain behaviors may involve various psychological modules. In other cases, a single

56 Hagen (2005), p. 146.

57 Downes (2010).

58 See for instance Moll et al. (2005).

59 While many evolutionary psychologists agree that the human mind is modular to *some* extent, they disagree about the extent and the functional details of this modularity. Proponents of the so-called massive-modularity hypothesis hold that EPMs are *domain specific*, i.e., tailored to particular problems. Others think that while mechanisms dealing with information uptake or output reactions might be domain specific, other mechanisms may be more general. E.g., the 'library model of cognition' illustrates how domain-specific output could result from *domain-general* information processing using domain-specific information: A domain-general information gathering mechanism collects information from domain-specific books in a library. The available evidence is insufficient to mark any of the alternatives as correct. However, the following discussion does not depend on the truth of either massive-modularity hypothesis or the library model. See Mallon (2008).

module (like a pleiotropic gene) might affect multiple behavioral traits (which makes them subject to stringent selection constraints).[60]

Each EPM has a three-part structure: It produces *output* by processing specific *input* in accordance with *decision rules* of an if-then form.[61] Input can be sensory data or information from other psychological mechanisms, output can take the form of physiological activity (e.g., arousal), information forwarded to other mechanisms, or behavior. EPMs interact: There is no *general* 'information encapsulation'[62], rather, information may run through several mechanisms that involve further sorting and specific if-then rules. Behavior can be a response to complex sets of conditions, where each is checked by a separate evolved psychological submechanism. The algorithmic or computational character of the psychological mechanisms in evolutionary psychology is a feature adopted from cognitive psychology and cognitive science.[63] It is important to keep in mind that even though their output was generally successful in the EEA, EPMs, like all adaptations, need not produce fitness-increasing behavior in *every* instance of their activation, even less if the environment has changed significantly.

Before I proceed, a methodological remark is in order: Advocates of an evolutionary understanding of the mind invoke evidence from various disciplines (e.g., behavioral studies, experiments on patients with localized brain lesions, imaging evidence) to support their views. Correspondence between mental phenomena and neural structures is a central building block of that paradigm. Hence, evolutionary psychology frequently involves hypotheses about neural correlates of mental phenomena (conscious and unconscious). However, the extent to which these hypotheses can be corroborated is subject to the current limitations of temporal and spatial resolution in brain imaging methods. Even though advanced imaging techniques detect activity changes in a cubic millimeter of brain tissue, the same level of activity can result from very different processes in the roughly 200.000.000 connections between the approximately 50.000 neurons in that space.[64] Experiments with patients that

---

60 See Mascaro et al. (2010), p. 39 and Churchland (2011), p. 97. These relations also depend on the degrees of freedom in the delineation of separate modules. EPMs can be examined at various levels of detail. Just as one can consider the liver as a functional unit, analyze the specific role liver cells fulfill as parts of that organ, or examine different structures within each cell, one might talk about an evolved psychological mechanism that generates aggression towards sexual rivals on a behavioral level or investigate the various subordinate mechanisms required for the respective behavior (such as uptake of visual information, evaluation of the opponents size, spatial orientation, activation of motor functions, signaling to regulate endocrine glands, etc.).

61 See Buss (2008), pp. 50–53.

62 This expression denotes the idea that EPMs do *not* 'talk to each other'. See ibid., p. 57, Hagen (2005).

63 See Downes (2010).

64 See Hagen (2005), pp. 154–155.

suffer from localized brain lesions can indicate which areas of the brain are necessary to perform certain functions. However, this kind of evidence by itself is insufficient to identify the neural components of any evolved psychological mechanism precisely. If a patient loses a particular ability, for instance, due to a stroke, and the damaged tissue can be localized, researchers can only infer that the process in question somehow implicates the damaged region, and that processes that remain unaffected do not necessarily recruit this area. The tissue could pertain to an $EPM_2$ upstream of an $EPM_1$ that is more immediately responsible for the ability in question, so that $EPM_1$ would operate if direct stimulation were to substitute input usually provided by $EPM_2$. For the project at hand, however, these limitations are inconsequential. They do not diminish the plausibility of the hypothesis that there are multiple EPMs designed by natural selection to deal with different adaptive problems, and that these mental mechanisms have (partly) separate neural substrates.

Critics claim that evolutionary psychological explanations are 'just-so stories', speculations about primeval living conditions of merely superficial plausibility based on an evolutionary perspective and whatever anecdotal information (i.e., archeological) is available. However, we know some of the circumstances under which our predecessors tried to survive and reproduce that allow for interesting hypotheses. For instance, several models in evolutionary psychology whose predictions have found empirical support were derived mainly from the fact that women, not men, give birth.[65] A related critical note points to the fact that experimental methods in evolutionary psychology are mostly those of 'traditional' psychology. This suggests that what is being tested is the presence of some mechanism or phenomenon, but not its evolutionary origin. This is correct if hypotheses are seen in isolation. However, if thinking about the human psyche in terms of the adaptive problems our ancestors faced frequently generates adequate predictions about psychological processes, that fruitfulness supports the evolutionary approach.

## 1.3 Emotion, Intuition, and the Situation Affect Moral Judgment

Recent empirical findings about moral judgment have sparked a debate about whether ordinary moral judgments are systematically error-prone. It is my understanding that this controversy is, beyond the impact of the bare observations, to a significant part a consequence of the fact that the seminal studies have interpreted these results in an evolutionary-psychological spirit. This chapter presents several major findings: Emotional intuitions sometimes determine moral judgments and behavior. These intuitive judgments can be surprisingly

[65] See Buss (2008), pp. 44–46.

insensitive to rational argument. Moreover, we regularly underestimate the influence of the situation. Important philosophical reactions to these findings, presented in chapter 2, express the concern that these influences are morally irrelevant.

Do these results in fact debunk the *inadequacy* of many day-to-day moral judgments and decisions? Those who attempt to demonstrate the irrelevance of specific determining factors, or even that granting normative authority to moral intuitions is always a mistake, frequently refer to evolutionary explanations of why intuitions respond to these factors. I will not address claims that moral *irrelevance* of certain influences is the main reason why these judgments are defective right away. Instead, I first attempt to establish how moral *relevance* could be spelled out in psychological terms (chapters 3 - 5) and assess the worries aired about particular influences on that basis. I begin by presenting seminal studies, which indicate that moral judgments are determined by evolved emotional, intuitive responses (1.3.1) and sometimes immune to moral reasoning (1.3.2). These findings spurred reactions critical of (evolved) intuition in moral judgment and the reliability of moral judgments more generally. Critics also draw on earlier research pointing to a certain corruptibility of moral judgment (chapter 1.3.3). Once I have brought out the suspicions of moral irrelevance, my investigation of the psychology of moral relevance begins with a look at the moral/conventional distinction, an important element of the cognitive-developmental tradition that dominated moral psychology until recently. As this account turns out to be too rigid and limited in scope, section 3.2 introduces moral foundations theory (MFT), which construes the moral domain more broadly. According to MFT, evolved 'moral emotions' are essential to the impression of moral relevance.

### *1.3.1 Emotions Shape Moral Judgment*

In a seminal paper published in *Science*, philosopher Joshua Greene and his colleagues investigated brain activity in moral judgments using functional magnetic resonance imaging (fMRI).[66] More specifically, they monitored subjects assessing the appropriateness of an action in various dilemmas.[67] Some of these dilemmas were variants of the trolley problem, a thought experiment well known in moral philosophy.[68] One of these variants is the 'switch

[66] See Greene et al. (2001). Functional MRI exploits the difference in the magnetic properties of blood carrying oxygen, and blood that has transferred the oxygen it was carrying to cells. Since oxygen consumption of brain cells increases with level of activity, the magnetic difference correlates with levels of activity in a given brain area. See Churchland (2011), pp. 123–125.

[67] Whether all the vignettes really have the structure of dilemmas was subject to debate, see the discussion in Kahane & Shackel (2010) and Sauer (2012).

[68] See Foot (1967) and Thomson (1985).

dilemma'[69]: A trolley is out of control and about to kill five people on the track ahead. The driver can divert the trolley to a different track before it reaches the group if he hits a switch on the dashboard.[70] In that event, however, a person on that other track is run over and killed. Is it OK to hit the switch? The majority of subjects say it is.[71] The similar 'footbridge dilemma', in contrast, is judged quite differently. Again, a trolley is out of control and headed for five victims. In this scenario, the agent is standing on a footbridge crossing the tracks between the trolley and the group of five. Next to him is a large stranger. Pushing the stranger off the bridge and into the path of the trolley would slow down the trolley, save the five and kill the stranger. Is it appropriate to sacrifice the stranger? Most subjects say it is not.[72]

Researchers have known about the intriguing contrast between the judgments subjects pass on these scenarios for a while. What remained unclear was *why* the respective acts are evaluated differently. After all, in each case, the choice is to sacrifice one live in order to save five. Discussions of the subjects' responses have either proposed that *both* cases ought to be judged equally (either permissible or impermissible), or tried to formulate a principle which accounts for the difference, much like linguists try to identify subconscious rules of grammar governing intuitive sentence formation. One such suggestion is the doctrine of double effect (DDE), according to which it is permissible to bring about a morally bad outcome if that outcome was not intended, but merely a foreseen side effect of an action realizing a (more valuable) moral goal.[73] One procedure to decide whether an effect is intended or unintended is to ask whether the agent would still have performed the act, had she not thought that the effect in question would occur. Applying the DDE to switch and

---

69 Greene et al. (2001) referred to this scenario as the 'trolley dilemma'. However, a new terminology (including 'switch dilemma') is adopted in Greene et al. (2009), p. 364.

70 In some articles, 'switch dilemma' refers to a case in which the switch is located next to the tracks (also known as 'bystander dilemma'). What I present here is the version used in Greene et al.'s original experiment. See http://www.sciencemag.org/content/293/5537/2105/suppl/DC1. In order to draw attention to this difference, the variant in which the switch is located *within* the trolley is sometimes referred to as the 'trolley driver' variant.

71 Participants were asked whether they considered the action by which five are saved "appropriate" or "inappropriate". In their 2001 article, Greene et al.'s language is ambiguous as to whether they are concerned with judgments of obligation or permissibility. "Appropriate" seems to be equally ambiguous. See also Greene (2005b), p. 58.

72 See Greene et al. (2001), p. 2105. Originally, the term 'trolley problem' refers to the *contrast* in judgments regarding a case in which a bystander can hit a switch to divert the trolley versus a case in which a surgeon can kill a healthy individual in order to save five lives with the organs taken from that person. "Why is it that the bystander may turn his trolley, though the surgeon may not remove the young man's lungs, kidneys, and heart? Since *I* find it particularly puzzling that the bystander may turn his trolley, I am inclined to call this The Trolley Problem. Those who find it particularly puzzling that the surgeon may not operate are cordially invited to call it The Transplant Problem instead." Thomson (1985), p. 1401, emphasis in the original.

73 See Cushman et al. (2006), p. 1083.

footbridge dilemmas yields the observed pattern of evaluations: In the switch dilemma, the intention is to save five lives, killing the single person is merely a foreseen side effect. The agent would equally have diverted the trolley had nobody been on the second track; consequently, the DDE declares hitting the switch permissible. In the footbridge dilemma however, the death of the large stranger is *not* an unintended side effect, but tied more closely to the action that is crucial in saving the five.[74] Since pushing the stranger amounts to intending a reprehensible act (killing), it is *not* permissible. However, the DDE does not appear to be the unconscious principle behind all moral evaluations of trolley cases. In the so-called 'loop dilemma', a switch diverts the trolley to a bypass that returns to the original track. The trolley would still kill the five after returning to the original track, were it not for a large stranger on the tracks in the loop. His weight, subjects are told, slows down the trolley sufficiently to let the others escape.[75] In this scenario, the ratio of subjects who consider the sacrifice permissible is higher than in the footbridge dilemma.[76] Yet, according to the DDE, it is equally impermissible to intend the stranger's death and let his weight slow down the trolley in both cases. John Mikhail tested two variants of the loop dilemma which preserve the means/side-effect distinction (large man on loop [means case] vs. man on loop in front of heavy object [side-effect case]). Diverting the trolley in the looped *means* case was judged morally worse, but the difference was much smaller than between the original switch/footbridge scenarios.[77] Even though few subjects justify distinctions made between side-effect and means cases by reference to principles like the DDE, "[t]hese results [might] suggest that the DDE is an adequate descriptive account for at least some part of the moral distinction between the fat man [i.e., footbridge] and bystander cases."[78]

Greene et al., however, suspected that judgments differ between footbridge and switch dilemmas for another reason: Killing a human being 'impersonally' by hitting a switch might be less emotionally engaging than 'personally' pushing someone in front of the trolley.[79] Neuroscientific research has identified brain areas whose activation corresponds to emo-

74 This line of reasoning can be criticized by arguing that the stranger's death is not intended after all. The agent would presumably consider pushing the stranger in the trolley's path even more seriously if he knew that the stranger might survive. Moreover, if the 'weight' in question were not a human being, but an inanimate object, he would also push it. Such complications have been noted; see for instance Greene et al. (2009), p. 370.

75 See ibid., p. 367.

76 See ibid., pp. 366–368.

77 See Cushman et al. (2010), p. 55.

78 Ibid., pp. 55–56.

79 Greene et al. (2001), p. 2106 distinguish moral-personal dilemmas from moral-impersonal ones and non-moral 'control' dilemmas.

tional arousal. If their hypothesis is correct, some of those brain areas should show increased activity in subjects dealing with the footbridge dilemma compared to those judging the switch dilemma; and that is what Greene et al. found.[80] Moreover, they believed they had observed that the minor group of subjects who considered pushing the stranger off the bridge *permissible* took more time on average to decide than the majority who considered this impermissible, while both groups showed similarly increased activity in emotion-related brain areas. However, Greene has since declared these particular findings invalid due to how response times had originally been aggregated.[81] The hypothesis in the 2001 paper was that subjects who judge the push permissible have to make an effort to overcome their initial emotional discomfort at sacrificing an innocent stranger. Greene et al. therefore predicted that decisions would generally take longer if they are incongruent with the subject's initial emotional response.[82]

As suggested above, the trolley problems have a history of discussion preceding Greene's investigations. Advocates of consequentialism typically consider the sacrifice morally permissible or even obligatory (depending on the question) in both cases, since, *ceteris paribus*, one death constitutes a smaller loss of well-being than the death of five. Deontologists often claim that it is never permissible to balance human lives or to kill, prohibiting the sacrifice in either case. Yet others sense a morally relevant variation between the cases and believe they should be judged accordingly. If we assume that folk judgments in the trolley cases are moral intuitions in the sense of being "natural, untutored judgments"[83], the range of positions can be taken to illustrate the varying degrees of importance moral philosophers attribute to such intuitions in their theorizing. On a pronounced anti-intuitionist view (as represented, for instance, by Peter Singer), it is *not* the task of ethics to devise principles whose application yields evaluations that match moral intuitions. Instead, moral judgments should be derived from overarching principles, and intuitions that conflict with judgments thus generated ought to be corrected rather than accommodated. Such an approach is not exclusive to consequentialist ethics; deontologists can entertain similar views. The DDE example above, however, illustrates a very different approach, namely the search for principles of moral evaluation that emulate intuitions quite closely. Positions in between the extremes of attributing normative authority to either *all* or *no* moral intuitions hold that

80 Brain areas associated with emotional activation include, among others, the amygdala, anterior cingulate and prefrontal cortices. See Dalgleish (2004) and Greene & Haidt (2002), pp. 520–521 for overviews.
81 See Greene (2009), p. 582.
82 See Greene et al. (2001), p. 2106.
83 Greene (2010), p. 19.

'basic' intuitions set important moral standards, while others are irrelevant. Other positions endorse 'correspondence with intuitions' as a general goal for theory building, while every counterintuitive theory-based evaluation can in principle be compensated for by other theoretical virtues (there are no 'basic', inviolable intuitions).[84]

Authors like Singer, who hold that normative moral theory should not pay too much attention to moral intuitions, have invoked Greene et al.'s research in support of their position. Greene et al.'s results supposedly indicate that emotions systematically influence intuitive moral judgments, and it is not clear why the propensity of a situation to elicit emotions should be morally relevant. Greene's evolutionary-psychological account of where emotional salience comes from and why it causes differential judgment reinforces such doubts.[85] His explanation rests on the assumptions of evolutionary psychology introduced in chapter 1.2.2: Firstly, mental phenomena correspond to physical events in the brain; secondly, like other species-typical physical design features, the brain was shaped by evolutionary processes, in particular natural selection. Consequently, at least some features of our mental makeup exist because they fostered the reproduction of their carriers and their genetic relatives; i.e., they increased inclusive fitness. Greene draws on this line of thought to explain the discrimination between the switch- and the footbridge dilemma.[86] Greene et al. used the following criteria to distinguish personal from impersonal dilemmas: Actions in personal dilemmas 1) cause a threat that has been authored and not merely edited by the agent (me), 2) can reasonably be expected to cause serious bodily harm (hurt), 3) affect a particular individual or group of people (you). Impersonal dilemmas fail to meet at least one of these conditions.[87] In ancestral times, an inbuilt first-personal emotional inhibition to killing people who did not pose a threat might have increased inclusive fitness (e.g., by avoiding revenge if the killing fails or by the victim's relatives, or by rendering beneficial

[84] This categorization maps the self-perception of moral theories. I believe that almost no normative moral theories adhere to either extreme; they differ mainly with respect to the intuitions they consider relevant rather than in their general influenceability by intuitions.

[85] See Greene (2005a).

[86] See ibid., pp. 345–346.

[87] See Greene et al. (2001), pp. 2107–2108, note 9; Greene (2005a), p. 345. Shorthand me-hurt-you criterion.

relationships less vulnerable to occasional disputes).[88] However, the methods of harming others available during the Pleistocene were presumably mostly "up close and personal"[89]: There was no machinery, no switches to be pulled. Under these circumstances, emotional discomfort at killing someone 'personally' would have sufficed to defuse many potential 'disadvantageous-kill' situations. The increased activation Greene et al. observed in emotion-associated brain areas of subjects dealing with the footbridge dilemma might represent such emotional inhibition. The switch dilemma is comparatively 'impersonal' because it (arguably) fails to fulfill the 'me' criterion; accordingly, interaction with the prospective victim is presumably not of the kind to which the inhibition mechanism evolved to respond. Therefore, the mechanism fails to activate, other considerations (numbers of lives lost and saved) processed by a more general reasoning capacity take the lead and generate rather 'consequentialist' judgment.

Greene et al. quote the fact that brain areas associated with working memory were generally less active when the moral dilemma was personal rather than impersonal (partly even less active than in the baseline condition) as evidence that reasoning processes are more important in impersonal dilemmas (see Figure 1, p. 28).[90] In response to the discovery of various variations of the trolley dilemma in which differences in evaluation were not explicable by reference to the personal/impersonal distinction as tentatively formulated in Greene et al.'s 2001 paper (me-hurt-you), Greene and colleagues have attempted to define the 'personalness' of an action in more detail. They devised dilemmas designed to isolate the influences on moral evaluation of spatial proximity between agent and victim, physical contact, and the presence of 'personal force'. 'Personal force' is present if the force directly manipulating the victim is generated by the agent's muscles (as in pushing someone, but

88 Interestingly, Greene is not very explicit in laying out the evolutionary rationale for an aversion to harmful actions. He states that one "might suppose that the sorts of basic, interpersonal violence that threatened our ancestors back then will 'push our buttons' […]" Ibid., referring to the advantageousness of avoiding physical harm, and "these responses evolved as a means of regulating the behavior of creatures who are capable of intentionally harming one another, but whose survival depends on cooperation and individual restraint" Greene (2008b), p. 43, referring to the advantageousness of not stifling cooperation through violence. It might be interesting to investigate whether evaluations of the trolley problems differ contingent on whether the subjects are asked to *imagine themselves in the position of the agent* (as trolley driver, on the footbridge, etc.), which was the case in the 2001 paper, or asked to evaluate the *action of another person*. Possibly, different evolutionary rationales/EPMs affect the patterns of judgment in each case because they mirror distinct adaptive problems. It might be advantageous to judge other aggressors negatively if that is correlated with avoidance of dangerous contemporaries. However, there might have been no equally strong selective pressure to refrain from harmful acts oneself. Such a constellation would yield the prediction that pushing the large stranger off the bridge elicits less of a negative response if I have to evaluate my own (hypothetical) action, rather than somebody else's. (*Footbridge pole* in Greene et al. (2009) [supplementary material] asks for an evaluation of an action undertaken by 'Joe'.)

89 Greene et al. (2001), p. 2106.

90 See ibid., p. 2107.

unlike hitting a switch, firing a gun, or opening a trapdoor, in which the force directly affecting the victim is not the agent's muscular force).[91] They explained the observed differences in judgments largely by the presence or absence of personal force; dilemmas in which the victim is killed by personal force were judged much less acceptable.[92]

Somewhat independently of which *specific* characteristics of the iudicanda[93] will turn out to determine the emotional activation in the different dilemma scenarios; anti-intuitionists and others argue that a possible evolutionary explicability of the efficacy of these factors does not bestow any moral force upon them; rather, such explanations could debunk the factors as morally irrelevant.

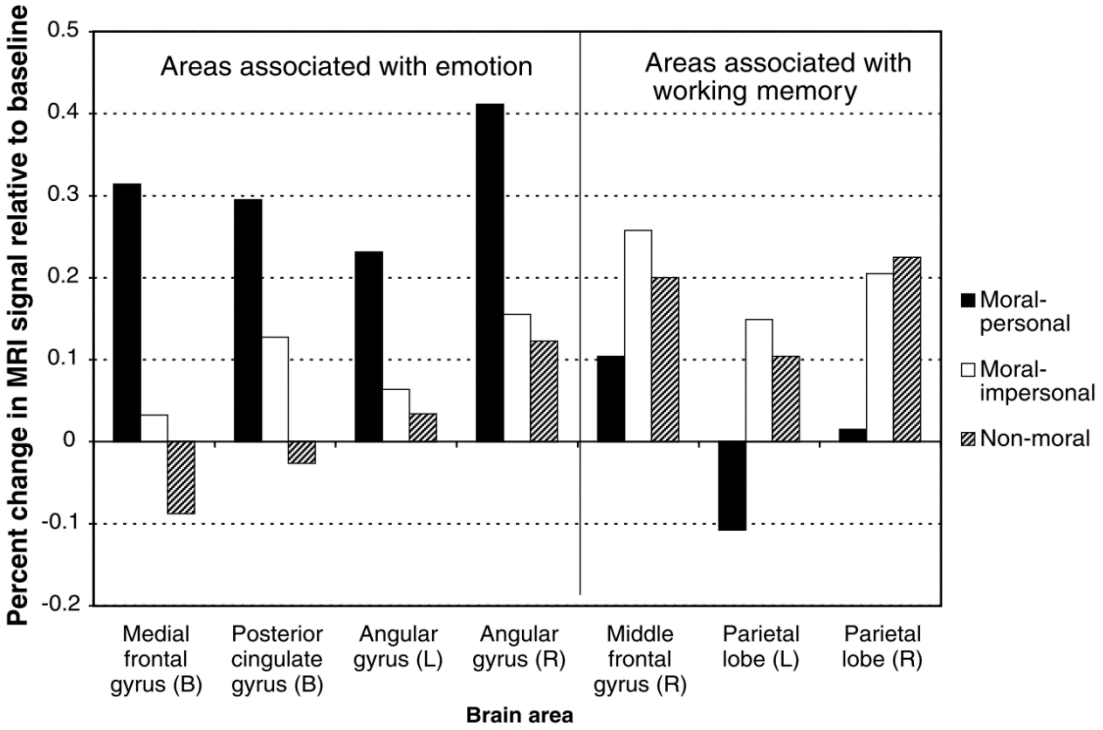


**Figure 1: Activation of Brain Areas in Three Types of Dilemmas**

From Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M. & Cohen, J. D. (2001): An fMRI Investigation of Emotional Engagement in Moral Judgment. In: *Science*, 293 (5537), pp. 2105–2108, p. 2106. Reprinted with permission from AAAS.

### *1.3.2 Emotion-Based Judgments Unaffected by Argument*

In addition to Greene et al.'s research, experiments conducted by social psychologist Jonathan Haidt and colleagues are often quoted as prime evidence for the importance of emotional intuitions in moral judgment. Haidt's results indicate not only that emotions play *some* role in moral judgment, but also that they *fully* control it surprisingly often. In such cases,

91 See Greene et al. (2009), p. 365.
92 See ibid., p. 369.
93 I will use this term to refer to whatever is being judged, be it actions, situations, or other matters.

moral judgments do not result from deliberation, nor do they respond to valid criticisms of potential justifications.[94] In several studies, Haidt confronted subjects with descriptions of "harmless yet offensive taboo violations"[95], asked them for their verdict on the transgression, and a justification.[96] Scenarios included, for instance, a woman cleaning a toilet bowl with the national flag, a family that cooks and eats its dog after it has been run over and killed by a car, and a man who purchases a dead chicken every week and uses it for masturbation before cooking and eating it. Perhaps most well-known, there is the story of Mark and Julie, an adult pair of siblings on vacation in France, who agree to have sex using two kinds of contraception, enjoy it, and afterwards decide never to repeat this experience and keep it a secret. Haidt constructed all scenarios so that the actions in question are not harmful. Nevertheless, many subjects, especially those of low socioeconomic status, judged the respective acts to be seriously wrong.[97] When asked *why* they condemned the act in question, subjects often cited some kind of harm (contracting disease from the dog or chicken carcass, a distorted relationship or handicapped children in the case of Mark and Julie, etc.), but held on to their judgment even after the experimenters had pointed out that the scenarios exclude those consequences. Haidt labeled this behavior 'moral dumbfounding', "the stubborn and puzzled maintenance of a moral judgment without supporting reasons."[98] Moral dumbfounding occurred *even* in subjects who acknowledged their inability to justify their judgment.

Observations of this sort led Haidt to propose that moral judgments frequently do not result from consciously weighing reasons for and against, but are instead expressions of quick, unreflected affective attitudes or 'moral intuitions'. As *psychologist*, Haidt uses the term 'intuition' in a narrower sense than what I described as its approximate meaning in *philosophical* contexts ('natural, untutored judgments'):

> *Moral intuition* is
>
> the sudden appearance in consciousness, or at the fringe of consciousness, of an evaluative feeling (like-dislike, good-bad) about the character or actions of a person, without any conscious awareness of having gone through steps of search, weighing evidence, or inferring a conclusion.

94 See Levy (2007), p. 292.
95 Haidt & Hersh (2001), p. 193.
96 See ibid., Haidt et al. (1993).
97 See ibid., p. 625.
98 Haidt & Hersh (2001), p. 193.

*Moral reasoning* is a

> conscious mental activity that consists of transforming given information about people (and situations) in order to reach a moral judgment. […] To say that moral reasoning is a conscious process means that the process is intentional, effortful, and controllable, and that the reasoner is aware that it is going on [99]

It is an essential feature of Haidt's psychological notion of intuitions that we do not have conscious access to the processes that produce them. This is not necessarily true of intuitions in the philosopher's sense (intuitions in the psychologist's sense are a proper subset of intuitions in the philosopher's sense).[100]

The reasons people gave appeared to be alibis devised to support whatever judgment they had previously passed. These findings fit well with others suggesting that the conscious self often automatically 'confabulates' an explanation for its behaviors, feelings, and judgments, because it is unaware of the mental *processes* that actually produce them. In these explanations, we draw on the mental *contents* that the conscious mind can in fact access: memories, current thoughts, or objects of attention.[101] If the reasons we give in order to explain our behavior or judgment to others and ourselves are not what really caused them, it is less surprising that arguments that undermine those confabulated reasons do not necessarily affect the behavior or judgment. Haidt, like Greene, proposes an evolutionary rationale for the influence of emotions on moral judgments. In short, it states that emotions evolved because they motivate us to behave in ways that solve adaptive problems, and that moral judgment springs from the activation of emotions that alert us to specific features of the iudicandum.

These results can seem disturbing. Surely, one might think, adequate moral judgment requires more than gut feelings; after all, we do not treat moral issues as mere matters of taste, but *argue* about them. What would be the point of moral argument if moral judgments were nothing but expressions of affective states on which reason and argument hardly have any influence? Moreover, is it not significant that some of these alleged 'confabulations' make more sense to us than others? We tend to conceive of moral attitudes as something that can be *justified* by reasons; what scientific accounts do is *explain* them. Supposedly, some moral judgments are right and others wrong because they relate to morally relevant facts in

---

99 Haidt & Kesebir (2010), p. 802. See also Greene (2010), p. 19 (page number refers to pdf document). Note that according to this definition, both moral intuition and moral reasoning are cognitive processes. Other authors use the term 'cognitive' to *exclude* uncontrolled processes.

100 See ibid.

101 See Wilson (2002), p. 97.

adequate or inadequate ways.[102] From this point of view, findings in the vein of those presented by Greene, Haidt, and their colleagues appear to uncover the *inadequacy* of at least some moral judgments. They are biased by morally irrelevant factors or weigh morally relevant factors wrongly, and ought to be substituted by more 'rational' or otherwise superior judgments that are immune to such mistakes. Moral evaluation should not depend on whether somebody is killed in a 'personal' or 'impersonal' manner; consensual incest should be judged on better reasons than a mere 'I just know it's wrong'.

### *1.3.3 The Power of the Situation*

There is a substantive body of research indicating that similar troubles affect the tendency to *act* morally[103]: Apparently, not only judgments, but also actions are quite sensitive to situational factors, and some of these factors are surprising. Such discoveries mark the so-called 'situationist' branch of social psychology, which challenges the claim that behavior is determined *mainly* by dispositions and character traits and holds that the situation is much more decisive than commonly suspected.[104] For instance, subjects in an experiment were almost five times less likely to help an injured man gather some books he dropped when there was a noisy lawnmower running nearby.[105] Being in a hurry, another factor which appears quite irrelevant to the presence of moral obligations, made subjects about six times less likely to offer assistance to a person slumped by the roadside,[106] while yet another experiment demonstrated that finding a dime in a phone booth exerted a large positive influence (factor 22) on the willingness to help.[107] The authors of this particular study suggest that success, feeling good, feeling bad, guilt, verbal contact, or the presence of other people also affect helping behavior.

Stanley Milgram's experiments on obedience showed how willingly subjects in the role of a teacher punished 'students' with dangerous electric shocks at the request of an experi-

---

102 See Darwall (1998), pp. 17–19.

103 Since the aforementioned experiments on moral judgment indicate that it is not quite clear what it means to act morally in specific situations, and what is morally relevant, I use "act with a moral dimension" in a hopefully uncontroversial minimal sense in this section: An act has a moral dimension if it affects the well-being of others. The definition is deliberately vague here; the delineation of the moral domain is the subject of chapter 3.

104 See Doris & Stich (2005), p. 118.

105 See Mathews & Canon (1975), p. 575.

106 See Darley & Batson (1973), p. 105. Some of the subjects were preparing to give a talk on the parable of the Good Samaritan, but that did not affect helping behavior significantly.

107 See Isen & Levin (1972), p. 387.

menter, indicating that obedience to authority can easily change or overrule the moral evaluation of an act.[108] Moreover, these experiments revealed a pattern of preparedness to punish quite similar to the pattern of moral evaluations in some of the trolley problems: The greater the psychological distance between teachers and their victim, the more readily punishment was executed. Obedience rates dropped from 65 to 40 percent when the teacher could see the victim; they increased to 100 percent when the student was in another room and had no possibility of communicating his distress, they dropped to 30 percent if the teacher himself had to make sure that the victim's wrists were placed on the electrodes.[109] Milgram even speculated about evolutionary origins of the propensity to obey, claiming that subordination is a crucial ability in societies with division of labor.[110] The similarly notorious Stanford Prison experiment documented how role-playing converted 'normal' college students (whose psychological inconspicuousness had been established beforehand) to cruel wardens and severely distressed inmates in a matter of six days.[111] Within a short period, some participants seemed to have radically adjusted their judgment of what kind of treatment of others was morally permissible, or had at least acted in a manner many would consider immoral.

The fact that these findings are surprising indicates that we frequently overestimate the influence of character and personality traits on behavior.[112] We cannot predict very accurately how a specific individual will behave in a particular situation by reference to personality alone. Overestimation of the predictive power of personality in combination with underestimation of the influence of situational variables on behavior has been termed the

---

108 See Milgram (1963).

109 See Smith et al. (2007), p. 822. Sixty-five percent obedience was the result in the standard condition, in which the teacher could hear the student, but not see him.

110 See ibid., p. 825. Milgram's results are frequently taken to illustrate how large parts of the German population came to participate in the atrocities committed by the Nazi regime. In fact, there are cultural differences in obedience rates: "Whereas American subjects were fully obedient 65 percent of the time, German subjects were fully obedient 85 percent of the time […]. A contrasting pattern was observed when a Milgram-style study was conducted in Australia: there, only 40 percent of the male subjects and 16 percent of the female subjects were fully obedient […]." Prinz (2007b), p. 279.

111 See Haney et al. (1973).

112 Some authors posit a 'predictability ceiling' that is "typically reflected in a maximum statistical correlation of .03 between measured individual differences on a given trait dimension and behavior in a novel situation that plausibly test [*sic*] that dimension." Ross & Nisbett (1991), p. 188.

*fundamental attribution error* (correspondence bias[113]), because it consists in mistakenly attributing behavior to personality rather than situational influences.[114] These results appear interesting because many believe that the efficacious situational factors identified *should* not make a difference, but also because we are usually unaware that they *do actually affect* moral judgment. In the next chapter, I take a closer look at some prominent positions regarding whether or not these findings expose the inadequacy of certain moral judgments or metaethical views.

[113] Appiah (2008), p. 42: "[The] tendency to ignore the role of context in determining behavior and to suppose that what people do is best explained by their traits rather than their circumstances […]".

[114] See Ross & Nisbett (1991), p. 189. Note also that the effects of personality and situation are typically intertwined: To some extent, persons chose the situations they put themselves in because of their personality traits, and they are chosen by others to handle particular kinds of situations because of these traits. Thus, the stability and reliability in the behavior of others is an effect of both the stability in their personalities as well as the similarity of the situations they typically encounter. Ibid. mention a clergyman and a criminal to illustrate this phenomenon: Each is perpetually placed in situations that promote behavior typical of a clergyman or a criminal, respectively. See ibid., p. 192.

# 2 Philosophers on the Significance of Moral Psychology

The beliefs about causes of moral judgment and action can be labeled more precisely than I have done so far. Commentators frequently argue based on 'moral relevance'. It is helpful to distinguish *moral relevance* from *efficaciousness in moral judgment*. This distinction relates to the notions of justification and explanation already touched upon, as well as the dichotomy of normative and descriptive ethics. Factors figure in justifications for actions and judgments because they appear morally relevant, while explanations of moral judgment and (im)moral behavior mention factors because they affect the explananda. Calling factor X morally relevant is a *normative* statement: It expresses the notion that factor X *ought* to be considered (in the right way) in order for a moral judgment to be adequate or for an action to be morally good. Stating that some factor X affects moral judgment or behavior is a *descriptive* claim about something that supposedly is the case, without evaluative component.[115] Both moral relevance and efficaciousness in moral judgment can be considered from descriptive and normative perspectives: *Descriptive* moral psychology collects data on what people consider morally relevant as well as on what actually determines moral judgment and action. *Normative* accounts make statements about what *actually is* morally relevant or about what people *should consider* morally relevant, as well as about what *should* actually *determine* judgment and action. Frequently, normative accounts will make similar recommendations on these subjects, but items of relevance and determinants of judgment may come apart because *considering* something relevant is a conscious process, while (arguably) not every factor that (should) affect judgment does so through consciousness. Which factors take effect subconsciously might be influenced by what people consciously consider morally relevant (also, since *ought implies can*, the considerable import of unconscious processes cannot be neglected by normative accounts). Moreover, there could, for instance on consequentialist accounts, be differences between what people *should* consider morally relevant and what actually *is* morally relevant. It might be the case that compatibility of an act with the Ten Commandments *should* be *considered* morally relevant, because it produces the best consequences in terms of harm (which is what *really is* morally relevant) overall. Apart from these combinations of perspective and phenomenon, the sets of factors which are considered morally relevant and those which actually affect moral judgment can probably have a partial overlap, which means that some factors that are considered relevant do not affect judgment, but

[115] The claim that something is a moral judgment or morally relevant behavior, however, can involve evaluations of moral relevance.

also that some factors which are not considered relevant do affect judgment. The research discussed in this dissertation also indicates that the set of factors considered relevant and the set of efficacious factors vary across cultures and individuals. Is there some core set of factors that is relevant and/or efficacious for all human beings? I will argue that answering such questions is complicated by the fact that there are various ways to describe the determinants of moral judgment (levels of explanation), but that the tendency of some factors to be universally relevant and efficacious can be understood from an evolutionary psychological perspective.

In the research discussed below, notions of moral relevance figure in two ways: On the one hand, they are the *object of investigation* in terms of the notions of moral relevance that specific populations of subjects hold and express in the justifications they offer for judgments and actions. In other words, notions of moral relevance are the subject matter of descriptive research efforts. On the other hand, we will encounter *normative* statements of moral relevance made by researchers (and others reacting to their findings) investigating what people consider morally relevant and what actually determines their moral judgment. They thus pronounce their views regarding whether some factor *should* affect moral judgment. These relevance judgments are frequently (yet not always, see for instance Greene's account, chapter 2.2) made without much more in terms of justification than an appeal to the 'obvious' moral irrelevance of some apparently efficacious factor.[116]

## 2.1 Singer Dismisses Judgments Owed to Evolution

Philosopher Peter Singer responds to the alleged discovery of morally irrelevant factors affecting moral judgment in *Ethics and Intuitions* (2005). He reviews Greene and Haidt's research and argues that emotional, intuitive moral judgments are defective.[117] While he agrees that evolutionary processes shaped important aspects of moral behavior, he considers this genealogy a reason why moral philosophers should *not* try to justify intuitive moral judgments by inventing principles to match.[118] Rather, they should stick to general principles (utilitarian principles, in Singer's case) and deduct moral evaluations from these. Singer claims that the evolutionary perspective can *explain* many aspects of morality, but not *justify* them. The idea that normative ethics should consider intuitive moral judgments is, in his

[116] Such relevance-judgments without argumentative support have been referred to as 'meta-ethical intuitions'. See Cushman et al. (2010), p. 67.
[117] See Singer (2005).
[118] See ibid., p. 348.

view, based on a fundamental misunderstanding of normative moral theory as methodologically analogous to science most prominently expressed in John Rawls's *reflective equilibrium*.[119] If a theory aims at *explaining why* we think about moral issues the way we do, then all moral judgments, including intuitive ones, belong to the body of data the theory has to accommodate. As in other empirical sciences, observations that appear incompatible with a theory cannot simply be dismissed, but stand to be either explained or rejected based on some plausible account of *why* they are erroneous. Yet, *explanation* is not the aim of *normative* ethics. Instead, according to Singer, it answers the question "*What ought we to do?*"[120] On this view, not all moral judgments are equally valid. Rather, some of them are more adequate than others are, and in order to find out what we ought to do, we need criteria that separate the former from the latter.[121] These criteria are provided by the most 'internally coherent and plausible' theory, independently of its correspondence with intuitive judgments.[122]

Once we appreciate the difference between normative and descriptive theory, and in the light of Greene's evolutionary explanation of why judgments in the footbridge and switch dilemmas differ, Singer urges us to agree:

> What is the moral salience of the fact that I have killed someone in a way that was possible a million years ago [footbridge dilemma], rather than in a way that became possible only two hundred years ago [switch dilemma]? I would answer: none.[123]

If factors without moral relevance ('salience') shape intuitive moral judgments, moral principles do not have to accommodate these judgments. In the strong version of Singer's view (see footnote 120), it is in fact *generally unnecessary* to adjust principles in order to match intuitive judgments, even if intuitions do respond to relevant factors. Rather, all principles follow from the overarching imperative to produce the best consequences (for instance, in

---

119 See ibid., p. 345.

120 Ibid. He adds: "It is perfectly possible to answer this question by saying: 'Ignore all our ordinary moral judgments, and do what will produce the best consequences.'" Ibid., pp. 345–346. More generally, he states, "[a] normative ethical theory […] is not trying to explain our common moral intuitions. It might reject all of them, and still be superior to other normative theories that better matched our moral judgments." Ibid., p. 345.

121 Henry Sidgwick, Singer's philosophical idol, proposed a related, threefold distinction between "[…] the psychological question, as to the existence of such moral judgments or apparent perceptions of moral qualities, […] the ethical question as to their validity, and […] what we may call the 'psychogonical' question as to their origin." Sidgwick (1874), p. 211. While I agree with Sidgwick that it is important to state exactly whether one is making psychological, ethical, or 'psychogonical' statements, I also believe that these concerns are closely intertwined. In particular, I will argue that how we deal with ethical questions is a psychological question, which can be answered in nonsuperficial ways only by reference to the origins of moral judgments or perceptions of moral properties.

122 See Singer (2005), p. 345.

123 Ibid., p. 348.

terms of preference satisfaction); if intuitive judgments respond to factors conducive or equivalent to maximization of preference satisfaction, so much the better for intuitions.

What is interesting in the present context is that it remains unclear exactly *why* Singer believes that the difference between 'being possible two hundred years ago' and 'being possible a million years ago' has no 'moral salience'; he seems confident that any reader will just agree that this is obvious. Moreover, he does not really argue for his even stronger claim that "there are *no* morally relevant differences between the two situations."[124] He merely states that "with our current powers of reasoning and our rapidly changing circumstances, we should be able to do better than that"[125] and that we should "attempt the ambitious task of separating those moral judgments that we owe to our evolutionary and cultural history, from those that have a rational basis."[126] Singer believes that 'more reasoned'[127] judgments are required to capture what is morally relevant. He dismisses the view that moral judgments essentially *are* intuitive (nonrational) responses, since it ostensibly leads to an unacceptable conclusion: moral skepticism.[128]

How does Singer know that evolved intuitive responses produce inadequate judgment? His reference to environmental change does not clarify that matter. Typically, environmental-change arguments in the context of evolutionary explanations point out that some design feature or behavioral tendency (e.g., a taste for sweet and fatty food) which was useful in a stone-age environment might be pernicious under altered circumstances (e.g., when sweet and fatty foods are not in short supply). However, what is the measure of this usefulness? In evolution by natural selection, only inclusive fitness matters: Being motivated to find sweet and fatty foods increased chances for survival when securing a sufficient intake of calories was a daily challenge. Today, the same urge is detrimental to health and physical attractiveness, and thus, to inclusive fitness. But why should inclusive fitness be morally relevant?[129] Singer might respond that he is not concerned with inclusive fitness, but instead with satisfaction of people's interests, and that a sweet tooth in today's environment poses a risk to the satisfaction of individual preferences. If concerns with preference satisfaction

124 Ibid., p. 350, my emphasis.
125 Ibid., p. 348.
126 Ibid., p. 351. Singer considers the intuition that "the death of one person is a lesser tragedy than the death of five" to be more rational than 'emotional intuitions'.
127 Ibid., p. 350.
128 See ibid., p. 351.
129 I do not want to adopt the unsatisfactory strategy of just pointing to the apparently obvious irrelevance of some concept or factor. I believe, however, that Singer should *argue* for its relevance if it were in fact at the base of his justificatory allusions. I do not believe that he considers inclusive fitness morally relevant. We will see later on why inclusive fitness is unlikely ever to be considered morally relevant.

were indeed the basis of Singer's dislike for moral intuitions, then his focus on the evolutionary history of these intuitions is misleading. Rather, he should explicitly discredit behavioral dispositions and moral intuitions, evolved or not, for their tendencies to obstruct preference satisfaction, as in: "Acting on the intuitive response to the footbridge dilemma results in suboptimal preference satisfaction." In that case, the evolutionary component is not necessary in the argument. Moreover, some dispositions with an evolutionary history might be detrimental, others conducive to preference satisfaction (think of useful adaptations like fear of heights). These considerations indicate that condemning intuitions because of their evolutionary history might not be a good idea. If it is at all possible to combine an evolutionary understanding of human nature with Singer's allusion to 'changes in the environment' to form a coherent position, it would certainly take much more than what is said in *Ethics and Intuitions*. On a related note, Singer's focus on the *evolutionary* history of some features of moral judgment might even be inconsistent if the (intuitive?) perception of people's interests as morally relevant is itself 'owed to evolutionary and cultural history'.

As for Singer's allusion to 'our current powers of reasoning'[130], he concedes that it is not clear what it means for a moral judgment to have a rational basis, apart from *not* being owed to our evolutionary and cultural history.[131] Consequently, it also remains unclear whether Singer considers the fact that a judgment has a 'rational basis' to be *sufficient* or merely *necessary* for that judgment to respond to morally relevant factors: On the one hand side, he does not spell out 'rational basis', so it remains unclear what exactly a rational basis achieves. On the other hand, however, he also does not hint at *further* conditions which, in combination with 'rational bases', would suffice for a judgment to have that desirable property. Because of this vagueness, I will not speculate further on what 'having a rational basis' could mean. A possible characterization of 'reasoned' judgments referring to specific psychological mechanisms figures in Greene's argument against deontology (chapter 2.2), others are offered in chapter 5.

Singer seems convinced that judgments owed to evolutionary or cultural history do not reliably respond to the morally relevant, or whatever else it is that establishes a 'rational basis' for judgment, *even though he is unable to explicate this crucial concept*. In fact, he even claims

[130] Although Singer does not elaborate on this point, our 'current powers of reasoning' can differ only in terms of cultural techniques from those available to the first humans with 'modern brains', for whom this brain structure was definitely an adaptation to the environment they actually lived in. This might be a problem for Singer's argument, since he also wants to dismiss moral judgments that we "owe to our cultural history".

[131] Note that Singer's position does *not* imply that moral judgments that we owe to our evolutionary or cultural history are *necessarily* inadequate. In case they are adequate, however, they are adequate only in virtue of their accidental correspondence with 'more reasoned' judgments.

that such judgments are *regularly* inadequate. This dissertation argues that most categories of value to which critics of evolved intuitions in moral matters resort are to some degree owed to evolutionary history as well, which means either that moral skepticism is correct, or that an evolutionary background does *not* generally render a moral judgment inadequate. Rather, even reasoned judgments (e.g., throwing the switch in the switch dilemma) probably depend on evolved aspects of human nature to an interesting degree. Consequently, conceiving of normative ethics as (even potentially) detached from moral intuitions appears increasingly less plausible the more we learn about the evolutionary background of these intuitions.

While Singer maintains that a better understanding of morality (better explanations of moral behavior) cannot have *immediate* normative consequences, it supposedly undermines conceptions of doing ethics that are too respectful of intuitions. Scientific investigations of morality can have indirect normative consequences by diminishing the trustworthiness of moral intuitions.[132] Greene's research is compatible with the idea that in some cases, reasoning overcomes initial emotional tendencies not to push the large stranger in the footbridge dilemma. Singer recommends these judgments since they 'properly' take account of the fact that one death is better than five deaths while disregarding the irrelevant 'technical' aspect of *how* the sole victim dies. They were also interpreted as evidence that we do not have to accept biased, emotional moral judgments.[133] Anticipating the objection that the attractiveness of the utilitarian position might just be the result of another intuition, Singer claims that if so, this intuition has no evolutionary background and is more 'rational'. Supposedly, rational utilitarian intuitions cannot have evolved since they express a love of 'mankind as such', but there was no interaction with 'mankind as such' in ancestral times. Moreover, he argues, it was hardly advantageous to be equally beneficent towards unrelated strangers and relatives alike.[134]

---

[132] See ibid., p. 349.

[133] Singer's argument builds on the reaction-time data in Greene et al. (2001). It seemed that 'utilitarian' judgments in the footbridge case took longer than 'impermissible'-judgments, this in turn was interpreted as an indication that an initial emotional disposition to judge pushing the stranger impermissible is 'overcome' by processes that are more rational. However, the claims regarding response times in ibid. have been abandoned by Greene. Greene et al. (2008) report that "[cognitive] load increased the average RT [reaction time] for utilitarian judgments by three quarters of a second, but did not increase average RT for nonutilitarian judgments at all." Ibid., p. 1151.

[134] See Singer (2005), p. 350.

## 2.2 Greene's Antideontological Argument

In *The Secret Joke of Kant's Soul* (2008), Joshua Greene considers the normative consequences of recent findings in moral psychology. In particular, he argues that

> […] our distinctively deontological moral intuitions (here, the ones that conflict with consequentialism) reflect the influence of morally irrelevant factors and are therefore unlikely to track the moral truth.[135]

While the alleged discovery of the influence of morally irrelevant factors on moral judgments plays an important role in both Singer's and Greene's papers, Greene specifically targets deontological ethics in his piece, underscoring an antideontological thrust already present in his dissertation (2002). In his view, consequentialist ethics are superior because they are not subject to the influence of irrelevant factors to the same degree. While strong, 'alarm-like' emotions drive deontology, consequentialist ethics crucially rely on different, more controlled mental processes. Since Greene attempts to name the mental processes which he takes to underlie consequentialism, his position could remedy an important shortcoming of Singer's account, namely the failure to provide an explanation of what marks 'more reasoned' moral judgments, and why they are supposedly more adequate than intuition-based judgments.

Greene's argument contains a descriptive part and a normative position. The *descriptive* part claims that consequentialism and deontology are "philosophical manifestations of two dissociable psychological patterns, two different ways of moral thinking"[136]: Deontological moral thinking is essentially an attempt at rationalizing domain-specific, emotional intuitions, while consequentialist moral thinking is the application of domain-general cognitive mechanisms to 'currency-like' emotional markers of moral relevance. Processes of the latter sort, Greene states, are closer to "genuine moral reasoning"[137] than the deontological mode of thought.

These claims employ specific definitions of deontology and consequentialism, as well as emotion and cognition. Greene defines deontological ethics as focused on rules, frequently expressed in terms of *rights* and *duties*, and involving intrinsic (nonconsequential) properties of an action, while the value of an action depends *exclusively* on its consequences in consequentialist ethics.[138] In a functional definition of deontology and consequentialism, he refers

[135] Greene (2008b), p. 70. Greene is an agnostic about the existence of moral truth Greene (2014b), p. 188), but is arguing here against a deontological position that assumes that moral truth exists.
[136] Greene (2008b), p. 37.
[137] Ibid., p. 36.
[138] See ibid., p. 37.

to differences in the judgments they engender. Judgments are 'characteristically consequentialist or deontological' depending on which position can more easily defend them.[139]

Greene uses the terms 'cognition' and 'cognitive' in a *narrow* sense that contrasts cognitive and emotional processes. It is important to be aware of this specific usage, since other authors (e.g., Jonathan Haidt) use 'cognition' more widely to refer to *all* information processing. In those cases, and in contrast to Greene's use, 'cognition' encompasses *both* emotional and nonemotional processes. Here is Greene's narrow definition:

> The rough idea is that 'cognitive' representations are inherently neutral representations, ones that do not automatically trigger particular behavioral responses or dispositions, while 'emotional' representations do have such automatic effects, and are therefore behaviorally valenced.[140]

Greene conceives of the relation between the two kinds of moral judgments and the two kinds of psychological processes as opposed to the orthodox allegiance to the primacy of either reason or emotion in moral judgment. Typically, deontological ethics in the wake of Kant is considered the rationalist alternative to the sentimentalist undertones of consequentialism present in the works of David Hume and Adam Smith.[141]

On Greene's picture, however, alarm-like emotions are essential to deontology, but not equally central to consequentialism. More specifically, he invokes Hume and assumes that the consequentialist weighing of benefits and costs does have an emotional component. However, it supposedly involves a different breed of emotions, namely *currency-like* emotions that convey *commensurable* value of a certain magnitude. *Alarm-like* emotions that motivate deontological judgments, on the other hand, convey nonnegotiable value that purports to dominate the decision at issue.[142] He does not intend the characterization of deontological to apply in every single case, but rather to capture how we *typically* arrive at such judgments. Thus, while the odd deontological judgment can result from cognitive application of the categorical imperative, Greene would consider his account adequate as long as *typical* deontological judgments *essentially* involve alarm-like emotional responses. 'Distinctively consequentialist' judgments (those which differ from deontological judgments), on the other hand, are not to be had without engaging in the cognitive process of weighing costs and benefits of an action.[143] To repeat: Typical consequentialist judgments *necessarily*

---

139 See ibid. It is not clear how 'ease of defense' is being measured.
140 Ibid., p. 40.
141 See Cushman et al. (2010), p. 54.
142 See Greene (2008b), pp. 64–65.
143 Ibid., p. 65.

*involve cognitive processes*, while typical deontological judgments do not (although they *sometimes* involve cognition). Typical deontological judgments are based on alarm-like emotions; typical consequentialist judgments involve currency-like emotions. Greene quotes various empirical findings in support of his descriptive hypothesis:

Greene et al.'s 2001 study (see chapter 1.3.1) and other publications on the neural correlates of moral decision making indicate that brain regions associated with emotion (and social cognition) are particularly active if what is being judged is a personal violation, and these in turn are the kind of violation that is likely to provoke characteristically deontological judgments. Impersonal violations, in contrast, supposedly do not trigger alarm-like emotions and thus make it easier to process the situation in a 'more cognitive' way.[144] Note that Greene's theory does *not* predict that personal violations *always* elicit deontological judgments. The actual outcome can depend on how eye-catching cost-benefit aspects are: The *crying baby*[145] and *infanticide*[146] vignettes both involve personal violations and thus, presumably, alarm-like emotional responses. However, subjects typically agree that it is wrong to kill the child in infanticide, but disagree about the evaluation of crying baby: Some people believe that it is permissible to smother the baby *in spite of* their alarm-like emotional responses to the personal violation involved. Greene et al. observed increased activity both in brain areas associated with 'response conflict'[147] and in areas associated with cognitive processes in *crying baby* as compared to *infanticide*. Among subjects who respond to the crying baby case, those who give a consequentialist answer (It is okay to smother the child!) show more signs of cognitive activity.[148]

Apart from neuroimaging evidence and response time data, Greene observes recurring patterns in moral debate and interprets the fact that these patterns can be reconstructed in terms of his theory as evidence in favor of it. For instance, he likens the structure of the argument surrounding Peter Singer's 'shallow-pond' example[149] to the anatomy of the trolley problem: Intuitively, the moral requirements differ between the situations that are being

144 See ibid., pp. 43–44.

145 "It is wartime, and you and some of your fellow villagers are hiding from enemy soldiers in a basement. Your baby starts to cry, and you cover your baby's mouth to block the sound. If you remove your hand, your baby will cry loudly, the soldiers will hear, and they will find you and the others and kill everyone they find, including you and your baby. If you do not remove your hand, your baby will smother to death. Is it okay to smother your baby to death in order to save yourself and the other villagers?" Ibid., p. 44.

146 "[A] teenage girl must decide whether to kill her unwanted newborn." Ibid., p. 45.

147 This term refers to situations in which two or more incompatible behaviors are triggered simultaneously. See Cushman et al. (2010), p. 51.

148 See Greene (2008b), p. 46.

149 In his well-known article *Famine, Affluence, and Morality*, Peter Singer compares the obligation of a single passer-by to rescue a child drowning in a shallow pond at the cost of a slight inconvenience to the obligation to relieve suffering far away. See Singer (1972).

contrasted (trolley/child in pond vs. footbridge/faraway needy), but finding a convincing account of *why* that should be so can seem difficult. According to Greene, the problem can be explained with reference to the distinction between personal and impersonal violations: While we evolved to have strong emotional responses to 'personal' interaction of the kind exemplified in the shallow-pond scenario, the same is not true for situations in which we can help via donations.[150] Judging it acceptable to spend money on luxury goods, instead of donating it, is characteristically deontological in the sense that it is more easily justified by reference to rights to spend one's own money however one likes than with reference to cost-benefit analysis.[151]

Further evidence comes from findings on the relation between the identifiability of victims and the willingness to help and the relation between anger and the willingness to punish. Both relations indicate that emotional responses predict when and how people will act in ways that are much more in line with deontological rather than consequentialist justifications. People are more willing to help identifiable victims even if an analysis of costs and benefits implies that the resources will be much more helpful if given to nonidentifiable victims. Moreover, the degree to which an action outraged those who are to decide predicts harshness of punishment much better than consequentialist considerations about the effects of punishment (deterrence, protection of the population from further harm done by the transgressor).[152] People appear to be indifferent to situational features that are very relevant to consequentialist justifications of punishment.[153] Even when they are asked to punish as a consequentialist would, they do not.

Emotional involvement also predicts condemnation of harmless actions (this was, for instance, a result of the studies by Haidt et al. quoted in chapter 1.3.2). Living in a Westernized culture, higher education, and increasing age appear to reduce the condemnation of harmless taboo violations, and are arguably indicators of a more cognitive approach to morality.[154] While it is true that deontologists do not necessarily condemn such actions, condemning judgments are more easily justified in deontological terms. Since the transgressions are designed to be 'harmless', they do not have (straightforward) harmful effects a consequentialist could refer to in order to justify a negative evaluation. Another study showed that, at least sometimes, emotion is sufficient to generate moral condemnation of a harmless

150 See Greene (2008b), pp. 46–47.
151 Voland & Voland (2014) consider the emergence of nonconsequentialist attitudes to be the fundamental explanatory challenge any account of the evolution of conscience has to meet. See ibid., p. 46.
152 See Greene (2008b), pp. 48–55.
153 See ibid., pp. 51–53.
154 See ibid., pp. 55–57.

action: Subjects were willing to condemn a student council representative who often tries to pick interesting topics for discussion when the action description contained a word towards which they had been hypnotized to feel disgust.[155]

Now why, according to Greene, do emotions and deontological judgments coincide? In his view, emotions are psychological mechanisms that spread due to their ability to motivate behavior which solved adaptive problems of social living (a detailed account of how they serve this function is given in chapter 4), and deontological philosophy is a "natural 'cognitive' interpretation"[156] of these emotions. In Greene's view, moral judgments thus spring from the output of EPMs. These EPMs are emotional rather than 'rational' because "emotions are very reliable, quick, and efficient responses to recurring situations, whereas reasoning is unreliable, slow, and inefficient in such contexts."[157] I would add that (at least some) emotional mechanisms stem from protoemotional motivational tendencies already present in our evolutionary ancestors. Natural selection builds on or modifies what is already present. It is less costly to recruit extant motivational mechanisms and shape them into an emotional system that deals with problems faced by humans living in groups than to develop new structures to that end. Rational, cognitive processes, in contrast, supposedly involve evolutionarily recent brain areas such as the neocortex.

Once the output of emotional processes enters consciousness, the human tendency to confabulate 'reasonable' explanations for these emotions kicks in, despite the fact that we have no access to the underlying unconscious processes.[158] The combination of these two traits, Greene argues, is the origin of deontological moral philosophy: Its judgments are in line with our alarm-like emotional intuitions, and the concept of a *right* is nothing but the conceptual manifestation of the experience that some actions (those violating the right) just *feel* wrong. Greene proceeds to defuse a possible problem for his claim that consequentialist ethics are superior: Why should *only* deontological ethics be confabulations? What about other moral philosophies?[159] In his view, consequentialism is inherently cognitive. Just as "there is a natural mapping between the content of deontological philosophy and the func-

[155] See Wheatley & Haidt (2005). This judgment is hard to justify in both deontological and consequentialist terms.

[156] Greene (2008b), p. 59.

[157] Ibid., p. 60.

[158] Greene draws on the moral-dumbfounding research mentioned in the presentation of Jonathan Haidt's arguments for the primacy of automatic processes in morality in chapter 1.3.2. See ibid., pp. 61–62.

[159] The view that *all* moral philosophy is a rationalization of moral emotions is, according to Greene, the strong form of Jonathan Haidt's position (ibid., p. 63). I disagree, since Haidt expressly mentions philosophers as exceptionally capable of reasoning. See Haidt (2001), p. 819.

tional properties of alarm-like emotions, […] there is a natural mapping between the content of consequentialist philosophy and the functional properties of 'cognitive' processes."[160] The spirit of consequentialism is "actuarial"[161]: It aggregates all aspects of the situation relevant to costs and benefits, can revise judgments in the light of additional data, and no aspect of an action has an a priori claim to determine the moral verdict. 'Cognitive' representations, such as the currency-like emotions that feature in consequentialist thinking, are less strongly valenced. Thus, there is more room for a balancing of concerns in consequentialist moral reasoning than in deontology, which is dominated by strong, rather inflexible (alarm-like) emotional responses to certain features of an iudicandum.

According to Greene, these descriptive claims have normative implications because they conflict with implicit factual assumptions in normative theories. In particular, science can illuminate where common moral intuitions come from, and moral intuitions of some sort are at the base of many or even most normative moral theories. To the extent that these theories rely on intuitive responses, they rest on the assumption that these responses are trustworthy. A better scientific understanding of the origins of intuitions, Greene believes, can undermine this trustworthiness.[162] Greene's normative argument has the following structure: He wants to show (descriptive claim 1) that deontological judgments are predicted by emotional responses to situations, and that, secondly, the predicting factor (emotion) is not systematically related to the concepts which deontologists *claim* shape their judgment (descriptive claim 2).[163] The normative claim is that the factor that *actually* predicts deontological judgments is morally irrelevant.

Thus far, I have outlined the argument supporting the first descriptive claim. In his discussion of the second empirical issue, namely the relation between the predicting factor and the concepts which supposedly determine moral evaluation according to deontologists, Greene limits the scope of his argument to what he calls 'rationalist deontology'. Rationalist deontological theories, like Kant's, ground judgment in "abstract theories of rights, duties, etc."[164] Because of their rationalist commitments, these deontologists cannot allow emotions any significance in the determination of moral evaluations. Nevertheless, the correlation between the judgments generated by these theories and emotional responses to the

[160] Greene (2008b), p. 63.
[161] Ibid., p. 64.
[162] He adds that his argument does not amount to deriving 'moral truths from scientific truths'. Ibid., p. 67.
[163] See ibid., pp. 67–68. This is a decisive move, because the moral relevance of the factors deontologists refer to is debatable, while those that Greene claims really predict their judgments are supposedly uncontroversially irrelevant.
[164] Ibid., p. 68.

respective situations is remarkable. Greene challenges the rationalist deontologists to explain how this is just a coincidence, and why emotions are not what determine many moral judgments. In his view, no convincing account of why emotions should correspond to the moral status of actions and situations is available. Greene's alternative explanation for the correspondence between deontological moral judgments and emotional response, on the other hand, is strengthened by empirical support both for the influence of automatic, intuitive processes on human behavior and for the tendency to rationalize the conscious output of processes to which we have no conscious access.[165]

In the normative part of his argument, Greene claims that rationalist deontologists, in order to defend themselves against the accusation of being influenced by nonrational factors, would have to provide a naturalistic account of how our emotional responses come to correspond to a somehow rationally (independently of emotions) discoverable moral truth. Apart from the empirical support for alternative explanations mentioned above, any such attempt faces an additional, fundamental difficulty: The factors that deontological judgments appear to respond to are *morally irrelevant*. This is an additional difficulty because rationalist deontologists not only have to explain how emotions coincidentally map rationally discoverable moral truth, but also how emotions can do this by responding to morally irrelevant factors. They would have to posit a systematic correlation between the occurrence of these morally irrelevant factors and morally relevant factors, something Greene considers highly unlikely. I quote Greene's illustration of the influence of morally irrelevant factors on deontological judgments in the footbridge case at length because I will later on claim that his argument is myopic:

> Take, for example, the *trolley* and *footbridge* cases. I have argued that we draw an intuitive moral distinction between these two cases because the moral violation in the *footbridge* case is "up close and personal" while the moral violation in the *trolley* case is not. Moreover, I have argued that we respond more emotionally to moral violations that are "up close and personal" because those are the sorts of moral violations that existed in the environment in which we evolved. In other words, I have argued that we have a characteristically deontological intuition regarding the *footbridge* case because of a contingent, nonmoral feature of our evolutionary history. Moreover, I have argued that the same "up close and personal" hypothesis makes sense of the puzzling intuitions surrounding Peter Singer's aid cases and the identifiable-victim effect, thus adding to its explanatory power.[166]

[165] See ibid., p. 69.
[166] Ibid., p. 70.

It seems to be the assumption that the characteristically deontological intuition in the footbridge case is owed to "a *contingent*, *nonmoral* feature of our evolutionary history" (my emphasis, BH) which makes it appear highly unlikely to Greene that the intuition reflects "deep, rationally discoverable moral truths"[167]. Intuitions, rather, are efficient instruments designed to solve adaptive problems in the EEA. Consequently, emotional intuitions can produce responses that do not *maximize* the balance of costs and benefits, independently of how costs and benefits are calculated.[168] Our evolved motivations do not operate directly on the concepts that matter in natural selection, like inclusive fitness. Nevertheless, they produced better results than alternative motivational dispositions in terms of inclusive fitness operating on the concepts they do operate on.[169] Greene then repeats the same argument for other typical conflicts between deontological and consequentialist judgments. Here is another version of what appears to be his core argument:

> [I]t seems that retributivist theories of punishment are just rationalizations for our retributivist feelings, and that these feelings only exist because of the morally irrelevant constraints placed on natural selection in designing creatures that behave in fitness-enhancing ways. In other words, the natural history of our retributivist dispositions makes it unlikely that they reflect any sort of deep moral truth.[170]

I quote Greene again because I want to make sure that we get this crucial step in his argument right. He refers to "contingent, nonmoral features of our evolutionary history" in the footbridge case and to "morally irrelevant constraints placed on natural selection". He claims that "our deontological intuitions […] reflect the influence of morally irrelevant factors"[171], speaks of "our moral emotions, which are sensitive to irrelevant factors"[172] and "intuitions [that] appear to have been shaped by morally irrelevant factors having to do with the constraints and circumstances of our evolutionary history."[173] So *what exactly* is nonmoral/morally irrelevant? At least two interpretations are compatible with the passages just quoted: Although it is not explicitly stated, Greene's references to the evolutionary process of natural selection could allude to the fact that the common denominator (and salient feature) of all evolved adaptations is their conduciveness to inclusive fitness in a particular environment. Nothing in this process seems to relate (systematically) to moral properties.

167 Ibid.
168 See ibid., p. 71.
169 We often evaluate neither hedonic costs and benefits, nor fitness-effects of an action, but merely respond to situations in ways that sufficed to generate fitness advantages.
170 Ibid.
171 Ibid., p. 70.
172 Greene (2008a), p. 117.
173 Greene (2008b), p. 75.

On this interpretation, the fundamentally *amoral* character of evolution by natural selection makes it unlikely that mechanisms produced by it track 'moral truth'.

The second thought that appears to be contained in the quotations is concerned with what our 'moral emotions' respond to in a narrower sense. Greene states that our moral emotions are sensitive to morally irrelevant factors. This can be read as a reference to the *actual* triggers of emotional responses, i.e., features of situations or actions (like 'personalness' of a situation). If this interpretation is correct, Greene is here ascribing moral irrelevance not to concepts like 'inclusive fitness' or the process of natural selection in general, but to more proximate, psychological causes of emotional activation.[174] Although Greene's essay is silent on the *relation* between the moral (ir)relevance of factors on these different levels of explanation, two ways to conceptualize it seem conceivable. On the one hand, and maybe this is closest to Greene's actual intention since he most frequently refers to the process of natural selection, the moral irrelevance of the concepts which do work in explanations that invoke natural selection is *transmitted* to the factors whose efficaciousness is a result of these processes. Call this *inherited* moral irrelevance: Moral irrelevance originates at a specific level of explanation and spreads to all more proximate explanations of the phenomenon under investigation. This conjecture is similar to arguments to the effect that moral properties cannot grow out of nonmoral properties. On the other hand, the moral irrelevance of concepts at one level of explanation might be *independent* of the moral irrelevance of concepts that figure on other levels. We might consider the concept 'inclusive fitness' and realize that it is not morally relevant just as we might consider 'personal force' and realize that it is not morally relevant, even if we do not know that 'personal force' affects moral judgment because of how natural selection works. A question to ponder is what kind of *psychological phenomenon* the evaluation of moral relevance on either level actually is. Is it an intuitive response, or is it the result of more cognitive analysis, evaluating the concept in question with respect to more or less well-formulated theoretical notions of moral relevance?

Remember that Singer, in *Ethics and Intuitions*, questioned the 'moral salience' of "the fact that I have killed someone in a way that was possible a million years ago [footbridge dilemma], rather than in a way that became possible only two hundred years ago". Maybe it was not his intention to have too much weight put on the exact wording, but for the sake

[174] In the context of biology, *proximate* explanations address 'how questions' and refer to the ontogenetic development and causal mechanisms that explain an individual organ's or trait's function. *Ultimate* explanations, in contrast, address 'why questions' and refer to phylogenetic development and adaptive function. These levels of analysis are also known as 'Tinbergen's four questions', named after the Dutch biologist.

of exploring the issue more thoroughly, I will assume that he really wants to advance this case by pointing to the moral irrelevance of the *difference in time* since a particular kind of action became feasible. I think this is more problematic than Greene's formulation because the interpretations of Greene's claim to irrelevance are more firmly anchored in the explanatory framework of evolutionary psychology: Both the procedural logic of natural selection as distal cause of behavior and the triggers of EPMs as more proximate causes of behavior do explanatory work within the research program. The amount of time that has passed since a type of action became possible, however, is neither a feature of a situation our intuitions evolved to respond to, nor a core explanatory concept in evolutionary theory. Nevertheless, Singer is concerned about the *evolutionary* history of certain judgments.[175]

Greene does not want to claim that responses which are (by-)products of evolution are *always* misguided, but rather, that it is unlikely that they correspond to an independently, rationally discoverable moral truth due to the way in which they came about.[176] Therefore, Greene's argument for skepticism towards (rationalist) deontological theories as a whole is, in his words, based on an *inductive* method: Given the available evidence, "the phenomenon of rationalist deontological philosophy is best explained as a rationalization of evolved emotional intuition."[177] He claims that this does not amount to deriving an 'ought' from an 'is'.[178]

Towards the end of the essay, Greene broadens the scope of his arguments to include less demanding forms of deontology (which do not aim to justify intuitions by rationalist theory) as well as other 'anthropocentric' approaches to morality. Anthropocentric moral philosophies, in Greene's terminology, take moral intuitions seriously. Greene argues that his argument for the moral irrelevance of some factors that determine these intuitions casts doubt on all of them.[179] Here, he seems to be very close to Singer in demanding general skepticism towards the role of intuitions in ethics. In these passages, he once again discusses Singer's shallow-pond case, and a third potential 'carrier' or origin of moral irrelevance emerges:

---

175 What do I mean by 'explanatory work'? It seems to me that we cannot give a proper evolutionary account of how some EPM came to be and how it works without referring to the notions of natural selection/inclusive fitness and the input that triggers the EPM. Timespan, on the other hand, is not as important. Rather, the significance of time in evolutionary explanations is a *result* of how the process works. Because it works by inheritance and because raising a new generation in a particular species takes a certain amount of time, evolutionary changes of a certain magnitude take a certain amount of time (depending also on the intensity of selection pressures).

176 See ibid., p. 72. However, it is not clear to me whether he wants to address the significance of something's *being a product of evolutionary processes* generally (by-product or adaptation) in these passages, or the significance of the specific fact that some response is a *by-product, rather than the actual adaptive response* which caused the evolutionary propagation of a particular psychological mechanism.

177 Ibid.

178 Ibid.

179 See ibid., pp. 74–75.

> […] [S]uppose that the *only* reason we say that it's wrong to abandon the drowning child but okay to ignore the needs of starving children overseas is that the former pushes our emotional buttons while the latter do not. And let us suppose further that the *only* reason that faraway children fail to push our emotional buttons is that we evolved in an environment in which it was impossible to interact with faraway individuals. Could we then stand by our commonsense intuitions?[180]

We can distinguish two levels of explanation here: In the first sentence, Greene refers to the activation of EPMs. He does not explicitly address the triggers of this activation, which have been subject to ascriptions of moral irrelevance before. In the second sentence, he refers to features of the EEA. Maybe this is a more sophisticated formulation of the point Singer was trying to make: The (*difference between the present and the) constitution* of the EEA of an adaptation and its by-products is not morally relevant. This kind of feature fits the earlier descriptions "constraints placed on natural selection" or "contingent, nonmoral feature[s] of our evolutionary history".

In sum, moral irrelevance is attributed to three types of concepts in Greene's text, the first two explicitly; the third one is my interpretation of several allusions:

1) *Proximate psychological causes* of moral evaluations, e.g., 'personalness' of an action
2) *Formative features of the EEA* of an EPM that affect moral judgment
   (i.e., more distal causes of the phenomena of type 1),
   e.g., the fact that it was impossible to interact with faraway individuals
3) *Fundamental factors* that structure evolution by natural selection, e.g., inclusive fitness, differential reproduction, inheritance, variability, scarcity of resources

The arguments brought forward by Singer and Greene are so-called debunking arguments. They claim to debunk the influence of morally irrelevant factors and thereby undermine the normative authority of moral philosophies that rely on the output of the psychological mechanisms affected. The fact that the efficacy of these factors can be explained with reference to evolutionary theory appears relevant in both Singer's and Greene's argument. However, it is not clear at which level of explanation or with respect to what sort of concept the charge of moral irrelevance first arises, how it is transmitted, and whether it arises at several levels independently. In personal communication, Greene explained that he thinks of the impressions of irrelevance as arising independently on several levels of explanation, by comparison of the factors identified to a list of items of moral relevance

[180] Ibid., p. 76.

that we know as a result of our socialization.[181] My goal is to evaluate the cogency of these arguments. To that end, I will investigate the notion of moral relevance more closely. Both Singer and Greene rely heavily on the ascription of moral irrelevance to various factors. I believe that we can view these attributions as psychological phenomena and explain them by reference to research that overlaps with the literature on which Greene rests his empirical claims. On the basis of a psychological account of moral relevance, I will then try to evaluate the strength of Singer's and Greene's arguments, whether they apply to normative moral philosophy that relies on intuitions generally (Singer), or specifically to those normative theories based on 'deontological' intuitions springing from alarm-like emotions (Greene). One of my suspicions is that the reference to evolutionary explicability in these debunking arguments is problematic, if not misguiding. Towards the very end of *The Secret Joke of Kant's Soul*, Greene mentions a problem that, I agree, is fundamental:

> Taking these arguments seriously, however, threatens to put us on a second slippery slope (in addition to the one leading to altruistic destitution[182]): How far can the empirical debunking of human moral nature go? If science tells me that I love my children more than other children only because they share my genes […], should I feel uneasy about loving them extra? If science tells me that I am nice to other people only because a disposition to be nice ultimately helped my ancestors spread their genes […], should I stop being nice to people? If I care about myself only because I am biologically programmed to carry my genes into the future, should I stop caring about myself? It seems that one who is unwilling to act on human tendencies that have amoral evolutionary causes is ultimately unwilling to be human. Where does one draw the line between correcting the nearsightedness of human moral nature and obliterating it completely? This, I believe, is among the most fundamental moral questions we face in an age of growing scientific self-knowledge, and I will not attempt to address it here. Elsewhere I argue that consequentialist principles, while not true, provide the best available standard for public decision making and for determining which aspects of human nature it is reasonable to try to change and which ones we would be wise to leave alone […].[183]

Does this passage not imply that the impression of irrelevance, which the factors at work in evolution by natural selection tend to elicit, is *not* decisive for the normative choice of which psychological processes we should rely on? I explore this issue and further questions regarding Greene's arguments below and in part III.

---

181 Personal communication, April 2012.
182 Giving away one's possessions until one is as badly off as the worst off.
183 Ibid., pp. 76–77.

## 2.3 Berker Disputes the Normative Significance of Neuroscience

In *The Normative Insignificance of Neuroscience*, philosopher Selim Berker offers a detailed critique of both Greene's empirical work on moral judgment and the normative conclusions drawn from it.[184] Moral irrelevance is a central notion in his paper, thus I will present his argument at some length.

The main point Berker attempts to make is that the normative significance of neuroscientific findings regarding, or evolutionary explanations of, moral judgment has been vastly overstated by Greene. In fact, he claims, all normative implications Greene presents are based on moral intuitions rather than implied by scientific results; all alleged implications moreover involve "shoddy" inferences.[185] At best, neuroscience can play a very indirect role in normative matters, namely by pointing to mechanisms involved in moral judgment that have proven unreliable in nonmoral contexts.[186] Berker considers the *structure* of the arguments made by Singer and Greene to be interesting, because it purports the potential to change the debate on first-order normative questions in a particular way in the face of scientific findings. The general pattern is to attribute certain (conflicting) moral judgments to particular psychological mechanisms and then argue that one (or several) of these mechanisms is unlikely to produce adequate judgments for specific reasons. The judgments produced by the incriminated mechanisms then appear less trustworthy compared to judgments produced by mechanisms not subject to such criticism. Berker proceeds to describe what he calls Greene et al.'s *dual-process hypothesis*:

There are emotional and cognitive mental processes, and there are characteristically deontological and consequentialist moral judgments. The hypothesis holds that

> characteristically deontological judgments are driven by emotional processes, whereas characteristically consequentialist judgments are driven by "cognitive" [i.e. nonemotional] processes, and these processes compete for one's overall moral verdict about a given case.[187]

A third distinction figures in Berker's recount of Greene's argument: The distinction between personal and impersonal moral dilemmas, formulated provisionally in terms of the

184 See Berker (2009).
185 See ibid., p. 294.
186 See ibid., p. 329.
187 Ibid., p. 301. This is in line with Greene's own characterization: "[C]haracteristically deontological judgments (e.g., disapproving of killing one person to save several others) are driven by automatic emotional responses, while characteristically utilitarian judgments (e.g., approving of killing one to save several others) are driven by controlled cognitive processes." Greene (2009), p. 581.

me-hurt-you criterion. This criterion was Greene et al.'s first attempt to spell out the features of a situation that give rise to characteristically consequentialist or deontological judgments in impersonal and personal dilemmas respectively. Berker then voices concerns about the empirical part of Greene's original argument:

Firstly, the connection between consequentialist judgments and nonemotional processing on the one hand and deontological judgments and emotional processes on the other hand might not be exclusive, for Greene's own data shows that (infrequent) consequentialist judgments in personal dilemmas correlate with neural activity in at least one brain region associated with emotion, namely the posterior cingulate.[188] In Berker's view, Greene reacts to this problem by admitting that consequentialist judgments also have an emotional component, although one which is grounded in currency-like rather than alarm-like emotional responses. Berker goes on to criticize this distinction as speculative and potentially question begging. I believe that this accusation rests on a misreading of Greene's account. The dual-process hypothesis, as I understand it, does not claim that there is *no* emotional activation in cases of consequentialist responses to personal dilemmas. Rather, the personalness of the situation elicits an alarm-like emotional response. If the subject nevertheless gives a consequentialist response (e.g., that it is OK to push the big stranger off the bridge, or to smother the baby), the alarm-like emotional response is being overridden by more cognitive (cognitive in the sense of 'neutral, not immediately action motivating') processes, which may involve currency-like emotional responses. Berker also cautions Greene against citing David Hume as an ally. In Berker's view, Hume did not only hold that (as he takes Greene to understand Hume) all *action* has an emotional component, but also that all *judgment* is in fact *driven* by emotions, so that there can be no conflict in decision making between reason and emotion, but rather only between different 'passions' or emotions. I believe, however, that Greene's account is compatible with Hume's: Since Greene admits for an emotional component in all judgments; his position does not imply a conflict between reason and emotion as such, but rather between less dominant emotions interacting with cognitive processes and emotions that in most cases do not allow for influence of cognitive processes.

Secondly, Berker points to problems in the way in which Greene et al. calculated the response time data in their 2001 article. Greene has conceded this point, but also points to later studies that avoid the problem. Since the compromised response-time data have not played a role in the argument thus far, I will not discuss the issue here.[189]

[188] See Berker (2009), pp. 307–308.
[189] See ibid., pp. 308–311, the same point made by McGuire et al. (2009), and a response in Greene (2009).

Thirdly, Berker claims that the personal/impersonal distinction spelled out in terms of me-hurt-you does not strictly correspond to the distinction between deontological and consequentialist judgments. In particular, a variant of the trolley dilemma called *Lazy Susan*[190] introduced by philosopher Frances Kamm counts as a personal moral dilemma according to the criterion, but regularly elicits typically consequentialist responses.[191] Greene agrees, and confirmed experimentally that this is a genuine counterexample to the hypothesis involving the me-hurt-you criterion.[192]

While these empirical issues are fascinating in their own right, Berker's reconstruction of Greene's normative argument is even more interesting. In combination with Greene's response to it, might provide some clarification as to how exactly Greene's claims of irrelevance are to be understood, and what significance the science of morality does or does not have. In trying to make sense of Singer's and Greene's arguments, Berker works through several possible 'bad arguments' for the conclusion that consequentialist judgments are superior to deontological judgments in order to finally address what he considers to be the strongest version of Greene's argument. Let me briefly mention these 'bad arguments', as well as the reasons for which Berker dismisses them:

A first understanding of Singer and Greene, the *Emotions Bad, Reasoning Good* argument, simply claims that while deontological judgments rest on emotions, consequentialist judgments do not.[193] In order to conclude that consequentialist judgments are superior to deontological judgments, a second premise stating that (and why) intuitions driven by emotions are unreliable is required. The same is true for Greene's distinction between alarm-like and currency-like emotional responses: Without an argument as to why the former are unreliable, deontological judgments cannot be dismissed. Moreover, the argument has to show why the processes underlying consequentialist judgments are reliable.

Secondly, Berker presents the *Argument from Heuristics*: It maintains that emotional processes drive deontological judgments, and that in nonmoral domains emotional processes tend to involve unreliable heuristics.[194] Thus, deontological intuitions are unreliable. There

---

190 "A runaway trolley is heading toward five innocent people who are seated on a giant lazy Susan. The only way to save the five people is to push the lazy Susan so that it swings the five out of the way; however, doing so will cause the lazy Susan to ram into an innocent bystander, killing him." Berker (2009), p. 311. Actually, the me-hurt-you criterion and personal force requirements are satisfied only if the lazy Susan is indeed pushed and rams into a bystander. If the lazy Susan just turns and thereby makes the train crash into a single person also sitting on the lazy Susan, then neither 'me' nor personal force are fulfilled, and a consequentialist judgment is thus no counterexample to either thesis.

191 See ibid., pp. 311–313.

192 See Greene (2008a), p. 108. *Lazy Susan* appears to be a counterexample to the personal-force criterion as well, since the agent's muscles do not produce the force that affects the victim.

193 See Berker (2009), pp. 316–317.

194 See ibid., pp. 317–318.

are at least two problems with this argument. Firstly, the notion of a heuristic typically implies that we have a grasp of the correct solution to a given problem, and can thereby evaluate to which extent the heuristic approximates the results of the optimal procedure. In the moral domain, however, it is often not clear what the right response to a problem is. Therefore, in Berker's view, the analogy is inappropriate. Moreover, it is uncertain that consequentialist intuitions do *not* also involve heuristics. Whether and to what extent moral intuitions are heuristics is the subject of chapter 9.4.

The *Argument from Evolutionary History* is a better representation of the allusions in Greene's and Singer's texts than the two aforementioned interpretations.[195] In a simple version, this argument claims that emotion-driven intuitive responses that correspond to deontological judgments are adapted to an environment we no longer live in, and that therefore these judgments have no normative authority. In Berker's view, this argument is problematic because consequentialist intuitions could also have evolved, and I agree. Greene claims that the evolutionary history of these intuitive responses makes it unlikely that they are correct, but the same dubious pedigree discredits intuitions involving currency-like emotions that drive consequentialist judgments.[196] At least, Berker writes, the argument from evolutionary history can be defended against the objection that the faculties we use to produce, for instance, scientific and mathematical knowledge have evolved as well, and that therefore the argument from evolutionary history would entail the implausible conclusion that scientific and mathematical judgments cannot be trusted either. While no obvious systematic relation obtains between moral properties and inclusive fitness, producing adequate representations of the environment was arguably a crucial feature through which the mental faculties involved in scientific or mathematical reasoning increased inclusive fitness, thus they are more likely to be truth tracking.[197] I might add that the development of methods of measurement that are more reliable than the natural human perceptual apparatus has been an important and successful scientific project for a long time. No comparable endeavor is underway in the field of ethics. On the other hand, I believe that the claim that intuitive moral-emotional responses have *no* relation whatsoever with moral properties is shortsighted and, in a way, puts the cart before the horse: Moral properties, I will argue, are *based on* such responses. There is probably no one-to-one mapping from emotional responses to moral properties. However, morality as we know it would not exist were it not for the evolved tendencies of human beings to respond to certain triggers by activation of

[195] See ibid., pp. 319–321.
[196] An evolutionary theory of consequentialist and other intuitions will be presented from chapter 3 onwards.
[197] See ibid., p. 320.

emotional psychological mechanisms (both alarm-like and currency-like, if one wants to adopt Greene's terminology). On such a view, morality is mind-dependent.

Meanwhile, Berker maintains that in order for the argument from evolutionary history to favor consequentialism, one would need to show why deontological intuitions are not truth tracking while consequentialist intuitions are, or at the least why they differ in *today's* environment, since the ultimate aim of the argument appears to be to discredit the *current* practice of deontological judgments.[198] Berker notes, however, that if the argument were to be thus extended, its normative force would come not from neuroscience, but from armchair theorizing about the relation between being truth tracking and being evolutionarily beneficial. As I will explain below, I do not believe that all considerations of this relation are armchair theorizing.

Finally, Berker presents what he takes to be the strongest interpretation of Greene's argument, the so-called *Argument from Morally Irrelevant Factors*. Since this argument is central to the following discussion, I quote it in its extended form:

| | |
|---|---|
| Premise 1 | The emotional processing that gives rise to deontological intuitions responds to factors that make a dilemma personal rather than impersonal. |
| Premise 2 | The factors that make a dilemma personal rather than impersonal are morally irrelevant. |
| Conclusion 1 | So, the emotional processing that gives rise to deontological intuitions responds to factors that are morally irrelevant. |
| Conclusion 2 | So, deontological intuitions, unlike consequentialist intuitions, do not have any genuine normative force.[199] |

Premise 2 is itself based on a normative intuition. This feature of the argument is, according to Berker, both a virtue and its crucial weakness.[200] It is a virtue because the appeal to an intuition about the irrelevance of whether a violation is personal or impersonal avoids the inadmissible inference of normative conclusions from purely scientific premises. On the

198 See ibid.

199 Ibid., p. 321. The unlike-consequentialist-intuitions part is not warranted by the premises. That would require a premise that states the factors consequentialist intuitions respond to, plus an assertion of their moral relevance.

200 See ibid., p. 322.

other hand, the recourse to an intuition supposedly shows that science (neuroscience, in particular) is 'normatively insignificant': Neither information about the *emotional* nature of the intuitions that allegedly shape deontological judgments, nor about the *evolutionary* history of these mechanisms is necessary for the *intuition* that the factors to which they respond are irrelevant.[201] You could just present people with the question "Is it morally relevant whether the death of the single victim in one of the trolley dilemmas has feature X?", where X is whatever feature researchers believe triggers the respective psychological processes, and they would have a presumably intuition-driven answer to the question. In combination with a claim that these processes drive deontological judgment, you have an argument against deontological judgment that does not involve neuroscientific findings or evolutionary explanations.[202] In Greene's particular case, Berker points out, the charge of irrelevance is leveled against the rather vague personal/impersonal distinction rather than the more explicit me-hurt-you criterion. Presumably, Greene expects more agreement to the claim that it is morally irrelevant whether harm is done in a personal or impersonal way than to the claim that "*whether one has initiated a new threat* [me] *that brings about serious bodily harm* [hurt] *to another individual* [you] is a morally irrelevant factor."[203]

While I have doubts regarding Berker's specific example, I do believe that it illustrates an aspect of the debate whose significance has not yet received sufficient attention. The concepts that figure in scientific accounts of moral phenomena differ in their aptitude to be processed by the mechanisms whose activation conveys to us the impression that something is a moral matter at all, or 'morally relevant'. I will provide further arguments for this claim in chapters 3 to 6. For now, I return to Berker. He sketches a 'best-case scenario' for the role of neuroscience in normative moral theorizing. Possibly, "[w]e notice that a portion of the brain which lights up whenever we make a certain sort of obvious, egregious error in mathematical or logical reasoning also lights up whenever we have a certain moral intuition."[204] If we can "see that the moral intuition in question rests on the same sort of confusion present in the mistaken bit of mathematical/logical reasoning, then of course we should discount the moral intuition."[205] He concludes that

[201] See ibid., p. 326. Note that Berker focuses on the irrelevance of what I have called *proximate* psychological causes of moral evaluations, rather than concepts on other, more distal levels of explanation.
[202] Depending on whether X contains neuroscientific or evolutionary-theoretical concepts.
[203] Ibid., p. 324, footnote 74. Emphasis in the original, insertions in brackets by BH. In my view, a more precise reconstruction of Greene's position would have to *contrast* ways of doing harm that fulfil the me-hurt-you criterion with ways that do not, and then ask for the moral relevance of the *difference* between these cases.
[204] Ibid., p. 329.
[205] Ibid.

> […] neuroscience can provide hints for where to look during our normative theorizing, but ultimately it can play no justificatory role in that task. Despite Greene's and Singer's claims to the contrary, learning about the neurophysiological bases of our moral intuitions does not give us good reason to privilege certain of those intuitions over others.[206]

## 2.4 Greene's Response to Berker and Further Statements

In a reply to Berker's article, Greene dismisses many of the methodological criticisms Berker raised and accuses him of incompetent and incomplete treatment of the relevant research. I will focus on his rejoinder to Berker's argument for the normative insignificance of neuroscience. While Greene agrees that deriving normative conclusions from scientific research requires normative premises, he thinks scientific results can nevertheless "do some work" in the argument.[207] Greene gives two examples: One is an argument for the conclusion that capital juries do not make fair decisions. It includes a normative premise (capital juries should regard the defendant's race as irrelevant) and a descriptive premise (capital juries' judgments are affected by the defendant's race). Since the conclusion could not be reached without the descriptive premise, Greene argues, it is obvious that the descriptive premise "does some work" in this normative argument.[208]

As a second example, Greene reports that scientific explanations of the intuitive aversion many feel with regard to incest can make people question their moral views on this issue. This is because science explains how the aversion could develop even though it is not 'true' or 'correct'.[209] He calls this kind of argument a debunking argument because it undermines values "by explaining their adoption in a way that makes it unnecessary or unlikely that those values are true or otherwise defensible."[210] A debunking argument does not show that there is no *other* justification for the evaluative attitude in question. If other justifications (justifications not debunked by the explanation) do *not* apply, the overall impact of the debunking argument increases. In some cases of consensual incest, plausible alternative justifications for a negative evaluation do not apply. (Remember the harmless taboo violations constructed by Haidt, chapter 1.3.2.)

Greene acknowledges Berker's 'argument from morally irrelevant factors' as an adequate rephrasing of his main argument.[211] Thus, it seems as if Greene's accusations of irrelevance

206 Ibid.
207 See Greene (2010), p. 7.
208 See ibid., p. 9.
209 See ibid., p. 11.
210 Ibid.
211 See ibid., p. 12.

actually aim for the *proximate* causes of moral intuitions, namely 'personal force', the most recent formulation of 'personalness'. Moreover, he tries to salvage the 'emotions bad, cognition good argument', the 'argument from heuristics' and the 'argument from evolutionary history' by presenting them as *supporting,* rather than *conclusive* evidence for the suspicion that judgments that are emotional, involve heuristic processes, and are products of our evolutionary history do not "reflect a sensitivity to factors that are morally relevant"[212]. The personalness of an act is *clearly* not morally relevant since this assessment is an intuition "that nearly all of us share, whether or not we have deontological or consequentialist proclivities."[213] Greene rejects Berker's worry that consequentialist judgments have similar problems. He believes that the cognitive, nonheuristic psychological processes underlying consequentialist judgments are less likely to go wrong. According to him, criticism against consequentialism[214] mainly refers to counterintuitive assessments it generates in specific cases. Because intuitions are so important in arguments against consequentialism, arguments that challenge the reliability of intuitions are, prima facie, a good thing for consequentialists. Deontological ethics, in contrast, are more often in line with intuitions (see the discussion of the DDE in chapter 1.3.1). In any case, (Greene considers this is a distinctive feature) deontologists cannot dismiss intuitions outright, while consequentialists can.[215] To argue that consequentialist judgments do not rely on intuitions, he distinguishes intuitions 'in the philosopher's sense' from intuitions 'in the psychologist's sense'. Philosophers use the term 'intuition' for *pretheoretical* judgments. For psychologists, in contrast, an intuition arises from information processing that is *not conscious*. All intuitions in the psychological sense are also intuitions in the philosophical sense, but not all intuitions in the philosophical sense are necessarily also intuitions in the psychological sense. According to Greene, consequentialists rely on intuitions in the philosopher's sense, but not on intuitions in the psychologist's sense (pretheoretical, but output of conscious processes). The unreliability he claims to have uncovered afflicts psychological intuitions, but not necessarily philosophical intuitions.[216]

Greene quotes empirical evidence to show that consequentialist judgments are not intuitive in the psychologist's sense:

[212] Ibid. He states that the relation between these characteristics of judgments and their sensitivity to morally relevant factors is "a contingent, probabilistic, and empirical one, not a logical one." Ibid.
[213] Ibid., p. 14.
[214] The "claim that aggregate consequences are the only thing that ultimately matters". Ibid., p. 18.
[215] See ibid., p. 20.
[216] See ibid., p. 19.

1) Conscious access: People can always give justifications for consequentialist judgments, but not for deontological judgments.

2) Lesion patients: Patients with damage to emotion-related brain areas make more consequentialist judgments.

3) fMRI data: Deontological disapproval is associated with emotion-related brain areas; consequentialist disapproval is associated with the DLPFC[217] (reasoning-related).

4) Priming counterintuitive behavior: If people are primed not to trust their intuitions, they make more consequentialist judgments, which seems to prove that these judgments appear counterintuitive.

Says Greene:

> One possibility–one that I favor—is that once all of the inner workings of our judgments are revealed by science, there will be nothing left for deontologists. All of the factors that push us away from consequentialism will, once brought into the light, turn out to be things that we will all regard as morally irrelevant. That's the grand ambition. The argument made here is just a first step.[218]

Why should we believe that deontological intuitions are wrong not just in this specific instance (trolley), but more generally suspicious? To make that point, Greene introduces the so-called camera analogy: According to this model, intuitions are the moral brain's automatic point-and-shoot settings. There is, however, also a 'manual mode'. It "includes our ability to apply explicit moral rules, to evaluate moral rules and judgments for consistency, and to override gut reactions that are at odds with our considered judgments."[219] Even though the automatic responses are always present, they can be overridden. Now the crucial question is in which cases one should rely on the automatic settings, and in which one should switch to manual mode. Greene states that the automatic settings, i.e., deontological intuitions, produce good results when applied to familiar problems, where familiar means similar in relevant aspects to the problems in response to which these automatic intuitions evolved. If the problem is unfamiliar in this sense, however, Greene argues that it is highly unlikely that a simple, automatic procedure produces a good response. This is true, Greene

[217] The dorsolateral prefrontal cortex, associated with higher-order cognitive functions.
[218] Ibid., p. 21.
[219] Ibid., p. 22.

argues, *regardless* of the criteria for what counts as a good response.[220] Greene also emphasizes that the dual-process hypothesis does not depend on the correctness of the personal/impersonal distinction, nor vice versa.[221] The relation between the two distinctions is rather a modular one: The personal/impersonal distinction is plugged into the dual-process hypothesis, which claims that dilemmas like the footbridge case elicit emotional-intuitive as well as cognitive responses; it is an attempt to spell out to which feature of dilemmas like the footbridge case emotional intuitions respond.

In sum, then, Greene seems to be making an argument against the application of deontological judgments based on emotional intuitions in fundamentally new situations. This argument rests on the suspicion that such intuitions respond to morally irrelevant factors, a suspicion which in turn rests on the assumptions that these moral intuitions were shaped by evolutionary processes, are of an emotional nature, and work as heuristics. All of these characteristics make it more likely that those intuitions respond to morally irrelevant factors, a suspicion confirmed, or so Greene thinks, in the case of the influence of personal force on moral judgment that his experiments identified.

In his book *Moral Tribes – Emotion, Reason, and the Gap between Us and Them*[222], Greene argues that our automatic, emotional psychological mechanisms are good at resolving conflicts of interest between the individual and the group it belongs to (me vs. us), but not good at resolving, and even responsible for, conflicts between different groups (us vs. them). This is because morality, in the shape of psychological adaptations, evolved not to enable universal cooperation, but rather cooperation within groups as a factor in intergroup competition.[223] Conflict is caused by "tribalism (group-level selfishness), disagreements over the proper terms of cooperation (individualism vs. collectivism), commitments to local 'proper nouns' (leaders, gods, holy books), a biased sense of fairness, and a biased perception of the facts."[224] Moral intuitions can be oversensitive (responding to morally irrelevant factors) and undersensitive (not responding to morally relevant factors).[225] Interestingly, Greene does not explicitly repeat the argument from morally irrelevant factors, but merely claims that factors (personal force) to which some moral intuitions respond are morally irrelevant.[226] In other places, his strategy is to show that distinctions like the one between

220 See ibid., p. 23.
221 See Ibid., p. 27.
222 See Greene (2014b).
223 See ibid., p. 26.
224 Ibid., p. 148.
225 See ibid., p. 212.
226 See ibid., p. 217.

means and side effects, or doing and allowing, which are morally relevant according to deontologists, but not utilitarians, have no moral authority because they are owed to "more basic cognitive mechanisms, ones that have nothing to do with morality per se."[227] He compares the 'footbridge switch' (in which the switch operates a trapdoor in the footbridge) case and the 'footbridge pole' case (in which the individual on the bridge is pushed with a pole) and argues that the physical mechanism by which the victim is killed is not in itself morally relevant.[228] From this we can conclude either that in both cases the action should be judged impermissible (in which case intuitions in the footbridge switch case are undersensitive), or else that the act should be judged permissible in both cases (oversensitive intuitions in footbridge pole).[229] Taking a closer look at judgment in different trolley cases, Greene finds that *both* the presence of what he calls 'personal force' *and* the distinction between killing as a means to an end and killing as a side effect determine our judgments. He considers the responsiveness to personal force to be an oversensitivity, a reaction to a morally irrelevant factor.[230] What about the means/side-effect distinction? Greene argues that we draw this distinction because of specific features of how we represent action plans. He assumes that humans developed a kind of intuitive reluctance to violent behavior when they became able to plan actions. From that point onwards, violence against other humans always carried the risk of revenge by the attacked or his relatives or friends, independent of their physical strength. This is because humans, as opposed to other animals, are able to kill their conspecifics using tools or while the enemy is asleep.[231] However, the action analysis process that identifies acts of violence is very simple and thus blind to side effects. Hence, Greene concludes

> […] the intuitive moral distinction we draw between harm caused as a means and harm caused as a side effect may be nothing more than a cognitive accident, a by-product. Harms caused as a means push our moral-emotional buttons not because they are objectively worse but because the alarm system that keeps us from being casually violent lacks the cognitive capacity to keep track of side effects.[232]

He makes a similar argument against the moral relevance of the distinction between doing and allowing. According to Greene, omissions are more abstract than actions, which is not what our brain evolved to deal with. Our cognitive apparatus is designed to deal with the

[227] Ibid., p. 241.
[228] In these passages, the argument is similar to the one outlined by Kumar & Campbell (2012).
[229] See Greene (2014b), p. 217.
[230] See ibid., pp. 216–217.
[231] See ibid., pp. 225–226.
[232] Ibid., pp. 239–240.

comparatively small number of things we *do* rather than the infinite number of things we do *not* do. Therefore, harmful actions are more emotionally salient than harmful omissions. Crucially, this difference between the mental representations of actions and omissions "has nothing to do with morality"[233]; therefore, "[this] hallowed moral distinction may simply be a cognitive by-product."[234] Summarizing these findings, Greene states that we respond to harm that is specifically intended (means), that we cause actively (doing), and by use of personal force. These three factors interact to make us respond negatively to acts that are 'prototypically violent', without sensitivity to potential benefits of the act.[235] This emotional alarm system is generally a good thing to have, but fallible. Harm caused as a means, actively, or by application of personal force, is not inherently worse than harm caused as a side effect, passively, or without personal force. These acts just *feel* worse because they tend to push our emotional buttons. The footbridge case triggers these negative intuitive responses, but because this response results from features of our moral cognition that have nothing to do with morality, we should not conclude from the majority's negative response in footbridge (which is unusual in that a violent action promotes the greater good) that it is sometimes *wrong to maximize happiness*.[236] Greene also identifies further factors that influence judgment but are, in his view, morally irrelevant. When deciding whether to help others, physical distance makes a difference: People believe that their obligation to help is greater if the one in need of assistance is closer.[237] Moreover, people are more willing to help victims of misfortune if they are identifiable at least in some minimal sense (for instance as victim #4) than if they are completely unknown.[238] A similar effect occurs with respect to punishment: People are more willing to punish identifiable transgressors, and their taste for retribution is generally insensitive to factors that matter from a consequentialist perspective, such as the probability of detection.[239] Instead, the willingness to punish seems to depend on the degree of emotional arousal. This is true even if subjects have to judge acts that occur in a deterministic universe, within which, when asked in the abstract, they claim individuals are *not* morally responsible for their deeds.[240]

In a recent article published in *Ethics*, Greene distinguishes a direct and an indirect route along which experimental research on the determinants of intuitive moral judgment can

233 Ibid., p. 245.
234 Ibid.
235 See ibid., pp. 246–248.
236 See ibid., pp. 250–251.
237 See ibid., pp. 260–261.
238 See ibid., p. 263.
239 See ibid., pp. 272–273.
240 See ibid., pp. 273–274.

have normative implications.[241] On the direct route, experimental information regarding the determinants of moral judgment is combined with "independent normative assumptions concerning the kinds of things to which our judgments ought to be sensitive"[242] to yield normative conclusions. In particular, moral judgments that are sensitive to morally irrelevant factors are inadequate moral judgments. Interestingly, experimental evidence can not only help us apply values (notions of moral relevance) we *already have* by making us aware of the influence of morally irrelevant factors, but it can also *change* people's values (such as a prohibition of consensual adult sibling incest) by providing a debunking explanation for the value in question.[243] The indirect route rests on the assumption that automatic mental processes can produce good results only if they have been shaped by some kind of trial-and-error experience (evolutionary, cultural, or personal). If problems are unfamiliar, however, it is better to rely on controlled processing. In combination with this tenet, experimental research on the processes underlying moral judgment can provide guidance regarding the question whether change in the processing of moral questions is recommended or not.[244] Since changes in processing might result in different moral judgments, this is a second and less direct route to normative significance of experimental moral psychology.

## 2.5 Kahane on Evolutionary Debunking Arguments

In 2011, philosopher Guy Kahane published an article in *Nôus* on the status of evolutionary debunking arguments (EDAs), that is, "argument[s] that claim […] that the evolutionary origins of certain evaluative beliefs undermine their justification"[245] in normative ethics and metaethics.[246] This definition seems to fit at least some of what both Singer and Greene have written, and I believe it is helpful to discuss Kahane's article in order to evaluate the merit of their arguments. In Kahane's view, EDAs are problematic at best: One of their problems is that they presuppose moral objectivism, which is a contentious position; another is that, even if objectivism is true, it is unclear how the destructive force of EDAs can be limited to only some normative positions, which is how EDAs are usually employed in normative ethics. Instead, they might imply what Kahane calls 'global evaluative skepticism',

241 See Greene (2014a).
242 Ibid., p. 25.
243 See ibid., pp. 27–29.
244 See ibid., pp. 30–32.
245 Kahane (2011), p. 104.
246 Note that Kahane understands EDAs as epistemic arguments; they deal with beliefs and their justification. The arguments are *not* about metaphysical/ontological questions like whether values exist. See ibid.

a worry that relates directly to Greene's question "where does the debunking of human nature stop?"

Before we get to these issues, let me introduce the tools Kahane uses to analyze the structure of EDAs. Debunking arguments claim that a certain belief or class of beliefs results from a process that, due to its very nature, is unlikely to produce accurate beliefs about the matter at issue. Kahane refers to these processes as *off track processes*: They do not track the truth. Psychological debunking arguments, for instance, might claim that a certain belief came about through an episode of motivated cognition. Debunking arguments contain both a causal and an epistemic premise. The causal premise states that a belief B was caused by a process X; the epistemic premise claims that X is an off track process[247]. For the causal premise to generate the required force, it is not enough that an off track process is *one of many* causes of a belief, since a full causal explanation will often contain off track processes. The influence of the off track process has to be strong enough to dominate potentially truth-tracking processes of belief formation.[248] I will refer to this as the *dominance requirement* for off track processes in belief formation. There is a second condition that has to be met for the argument to work, namely the absence of alternative justifications which might support the belief post hoc, even if it came about through an off track process. Only if this condition is met can the argument attack the general *justification* of the belief in question. Call this the *no other justification condition*. In Kahane's presentation, the causal premise meets this requirement if explanation of the belief via the influence of off track processes by itself renders alternative (post hoc) justifications implausible.

The conclusion of the debunking argument states that B is unjustified, especially once the individual who holds belief B is aware of the off-track nature of X. Note that the argument does *not* establish that B is in fact false, but only that it is unjustified to believe that the content of B is the case. B might be right, but that would be an improbable coincidence if X were indeed an off track process and dominant in the formation of B. Neither does the argument establish that belief in ~B is justified (it is relatively more justified than belief B at most).[249]

Let us consider the general structure of evolutionary debunking arguments.

247 On Kahane's view, the epistemic premise does not require that we know what a truth-tracking process would look like. See ibid., p. 106.
248 See ibid.
249 See ibid., p. 108.

| | |
|---|---|
| Causal premise | Moral judgment Z is (indirectly) owed to evolutionary processes. |
| Epistemic premise | Evolved psychological mechanisms do not track what moral judgment is supposed to track. |
| Conclusion | Absent further justification, there is no reason to believe that moral judgment Z is adequate. |

The arguments of both Singer and Greene fill out Kahane's general scheme in a specific manner: They are, to be exact, evolutionary debunking arguments against intuitive evaluative beliefs. Their causal premise states that we have an evaluative belief P because we have an intuition P that is caused by evolution. The epistemic premise holds that evolution is an off-track process with respect to "evaluative truth"[250]. With respect to the plausibility of the causal premise, Kahane notes that what is required for the argument to work is just that *some* evolutionary explanation is more plausibly the origin of the evaluative belief than alternative, possibly truth-tracking processes (*requirement of dominance*).[251] Now, Kahane claims that such arguments "appear to presuppose the truth of *objectivism*."[252] In his view, objectivists

> claim that evaluative propositions have truth conditions that are not grounded in our evaluative attitudes; anti-objectivist views (which include a range of subjectivist, response-dependent and intersubjectivist views) deny this.[253]

If you believe that evaluative propositions have truth conditions, and that these are grounded in our evaluative attitudes, Kahane argues, it does not make sense to worry about whether the processes that produce these attitudes were or are truth tracking, because there is no evaluative truth other than the output of these processes.

Since moral objectivism is a contentious position, a commitment to it could considerably weaken the appeal of EDAs.[254] The question we have to answer in order to address this worry is whether nonobjectivist debunking is in fact inconceivable. Finally, Kahane argues

---

250 Ibid., p. 111.

251 See ibid. Note also that Kahane explicitly refers not only to natural selection, but *all* evolutionary processes (including genetic drift) as off-track processes. See ibid., pp. 111–112.

252 Ibid., p. 112.

253 Ibid., p. 121, endnote 1. Sharon Street's definition of *realism* about value is very similar: "The defining claim of realism about value, as I will be understanding it, is that there are at least some evaluative facts or truths that hold independently of all our evaluative attitudes." Street (2006), p. 110. She also refers to this position as "the view that there are *mind-independent* evaluative facts or truths." Ibid., p. 156 note 1.

254 See Kahane (2011), p. 113.

that EDAs threaten to undermine the justification of beliefs beyond the confines of the normative positions they target.[255] After all, the ability to care for others that motivates utilitarianism, but also selfish concerns, are themselves susceptible to evolutionary explanation.[256] Possibly, all evaluative beliefs are susceptible to such explanations. Do targeted EDAs in normative ethics spin out of control and force objectivists to adopt "global evaluative skepticism"[257]? In terms of the structure laid out above, global scope manifests in the extension of the causal premise and the conclusion to *all* moral judgments. In order to assess this conjecture, we have to explore how far the influence of evolutionary processes actually extends.

It seems quite plausible that the ability to care for the well-being of others is a result of evolutionary processes. Somewhat surprisingly, Kahane does not explicitly check for *dominance* and *absence of other justifications*. Even if, on a charitable interpretation, we assume that natural selection and other evolutionary processes satisfy the dominance condition, it is not obvious that there is no alternative justification for the tendency to care for others. If there were, the EDA against caring for others would not go through, according to Kahane's own criteria. Kahane goes on to discuss a metaethical argument put forward by philosopher Sharon Street that attempts to establish the kind of global evaluative skepticism he believes a stringent application of the EDAs used in normative ethics implies. According to Kahane's reconstruction, this argument has the same structure as targeted EDAs in normative ethics, but employs the more general causal premise that all (moral) evaluative beliefs are owed to evolutionary history.[258] In combination with the assumptions that 1) evolutionary processes do not track evaluative truth and that 2) objectivism is the correct account of evaluative properties, this yields *global evaluative skepticism*: None of our evaluative beliefs are justified.

Kahane questions the global causal premise. Given observable evaluative diversity, it seems unlikely that all evaluative beliefs lend themselves equally to evolutionary explanation. If evolutionary processes do not fulfill the dominance requirement for some evaluative beliefs, then skepticism regarding those beliefs is not warranted, but rather a lowering of justification in proportion to the aptitude of a specific belief to be explained in evolutionary terms (again, he does not discuss the no-alternative-justification requirement). However, even if *some* evaluative beliefs are not subject to straightforward evolutionary explanation,

255 See ibid., pp. 113–114.
256 A point also made by Berker (2009), p. 319.
257 Kahane (2011), p. 114.
258 See ibid., p. 115.

Kahane suspects that evolutionary explanations carry farther than many think.[259] In particular, utilitarian judgments are not immune to evolutionary debunking. Rather,

> [i]f any EDA is successful, an EDA of partial altruistic concern must be. But this means that extending this concern through reasoning does nothing to salvage its epistemic status.[260]

Moreover, utilitarianism requires an account of well-being. The beliefs that pain is bad and that pleasure is good are obvious candidates to figure in that account; they are also obvious candidates for evolutionary explanation. While at first glance it might seem adequate to ascribe the generic epistemic premise mentioned above to both Greene and Street, Kahane's account pushes important differences in the manner of presentation of the respective EDAs to the background. A closer look at Street's work reveals an argument against objectivism whose power depends on the pervasiveness of evolutionary influences on moral judgment; Greene's position crucially depends on intuitions about moral relevance, and thus illustrates what nonobjectivist debunking might look like.

## 2.6 Street's 'Darwinian Dilemma' for Objectivism

Sharon Street writes about realism rather than objectivism, but her definition of realism is very close to what Kahane defines as objectivism.[261] In her case, Kahane's epistemic premise is a corollary of the 'Darwinian Dilemma', which, according to her, proponents of mind-independent evaluative truth (objectivists) find themselves in if the effects of evolutionary processes indeed pervade our evaluative attitudes. In Street's view, realists have to take a position on the *relation* between the selective pressures that shaped our evaluative judgments (to some extent) and the supposedly mind-independent evaluative truths the realist posits. *Denying* any relation between the processes that shaped evaluative judgments and evaluative truth would imply that the psychological mechanisms that shape moral judgment to a significant degree produce adequate judgments merely as a matter of chance, or are mostly mistaken. To Street, this skeptical conclusion is so implausible as to be unacceptable.[262] Could the use of reflective reason enable us to eliminate these distorting evolutionary influences and thereby arrive at judgments that are not susceptible to evolutionary debunking? Street concedes that our reflective nature will make us respond to any perceived illegitimate

259 See ibid., p. 119.
260 Ibid., p. 120.
261 See footnote 253.
262 See Street (2006), pp. 121–122.

influence on our moral judgments. In her view, however, rational reflection depends on evaluative assessments that are influenced by evolutionary processes. It uses some of the evaluative perceptions shaped by evolution to assess other matters. Reflection cannot proceed without evaluative premises, and the arguments for the shape-giving influence of evolutionary processes on evaluative attitudes apply to this set of evaluative premises.[263] Thus, pointing to reflective reasoning does not help the realist escape the implausible skeptical consequences of denying any connection between evaluative truth and the processes that shaped our evaluative attitudes.

What happens if the realist *does* assume a relation between the workings of natural selection and mind-independent evaluative truth? Prima facie, this strategy has the advantage of not implying that most of our evaluative judgments are false. According to Street, the realist, because of her metaethical commitments, has to posit that the relation is a *tracking* relation: Evolutionary processes track evaluative truths, since tracking 'what one has reason to do' is evolutionarily advantageous. On such a tracking account, considering one's own survival and the survival of one's offspring valuable confers advantages in terms of inclusive fitness on those who have this attitude *because that attitude is true*. Street argues that the tracking account of the connection between evaluative truth and natural selection is inferior to an alternative explanation of why certain evaluative judgments conferred fitness advantages upon our ancestors in terms of parsimony, clarity, and explanatory power.[264] The alternative she proposes, an *adaptive-link account*, holds that some evaluative attitudes were conducive to inclusive fitness not because they tracked moral truth, but rather because they made our ancestors respond to their environment in ways that enhanced reproductive success.[265] For instance, a positive evaluation of a person that helps us motivates us to return the favor, thus enabling us to reap the fitness benefits of cooperation. In the case of evaluative judgments, Street writes,

> the link between circumstance and response is forged by our taking of the one thing to be a *reason* counting in favor of the other – that is, by the experience of normativity or value.[266]

[263] See ibid., pp. 123–124.
[264] See ibid., p. 129.
[265] See ibid., p. 127.
[266] Ibid., p. 128.

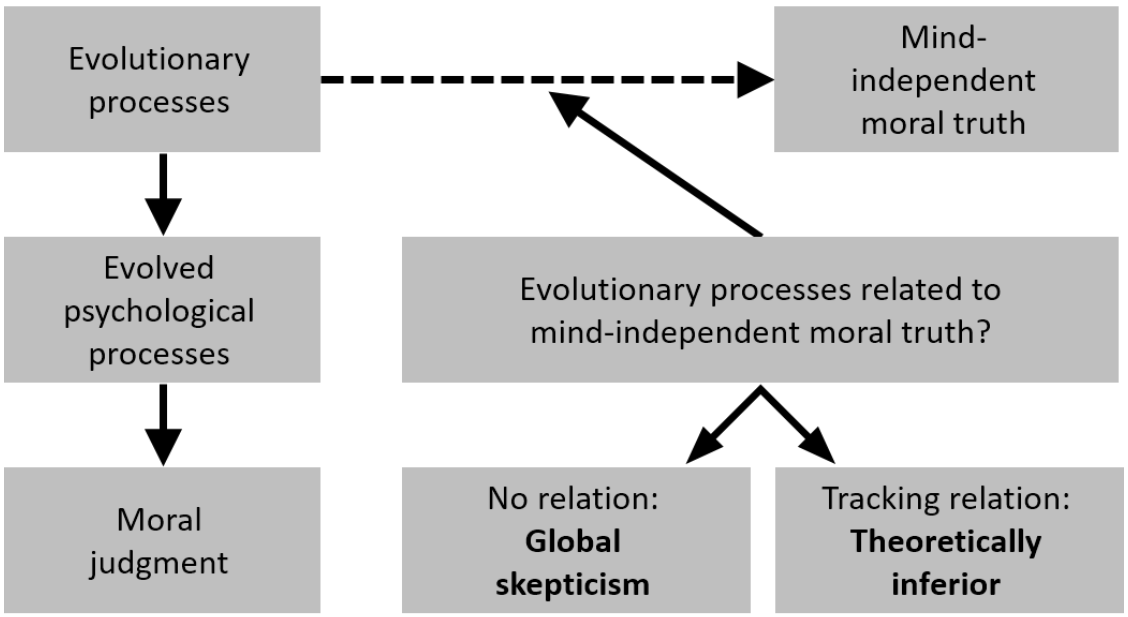


**Figure 2: Street's Darwinian Dilemma for Objectivism**

Illustration by BH

The adaptive-link account is more parsimonious because it does not posit mind-independent evaluative truth. The adaptive-link account is clearer than (a vague version of) the tracking account: It can explain in detail how the evaluative attitude in question contributes to inclusive fitness, rather than obscurely claim that recognizing evaluative truth contributes to inclusive fitness without reference to any mechanism by which this effect might come about.[267] This is particularly evident when the adaptive-link account is compared to *nonnaturalist* realist positions, since nonnatural properties are supposed to have no causal influence on the natural world and thus, *a fortiori*, not on reproductive success.[268] Finally, the adaptive-link account often can explain why human beings have some evaluative attitudes rather than others by reference to the fitness effects of certain behaviors in specific circumstances. In comparison, the tracking account merely states that ultimately, we make these judgments rather than others because they are true. It fails to explain why so many truths *align* with the judgments that the adaptive-link account can explain without reference to truth. On the other hand, it has difficulties explaining the frequent occurrence of false judgments: Given that many would agree that one should not discriminate between members of the in- and out-group in terms of rights and the like, the tracking account has difficulty explaining the well-documented human tendency for in-group favoritism.[269] Finally, the tracking account

[267] It is not the case that for every kind of truth, recognizing it is advantageous. Street mentions the detection of electromagnetic waves of the lowest frequencies as an example. See ibid., p. 130.
[268] See ibid., p. 131.
[269] See ibid., pp. 129–130.

does not advance our understanding of why certain evaluative judgments just do not occur with any noteworthy frequency (such as the judgment that plants are more valuable than human beings are).[270] The adaptive-link account, in contrast, can explain such judgmental tendencies by reference to their fitness effects. The adaptive-link account suggests a unifying feature that marks (many of) our evaluative attitudes, while the tracking account does no such thing. Street argues that these arguments suffice to judge the tracking account inferior. Crucially, Street argues, *all* accounts the realist can offer of the relation between the selective pressures that shaped our evaluative attitudes and independent evaluative truths *have to be* tracking accounts. In her view, those who believe in an independent evaluative truth cannot otherwise accept the finding that human evaluative attitudes are to a significant extent shaped by selective pressures and at the same time avoid the skeptical conclusion that the vast majority of our evaluative judgments is off track. If evolutionary processes do not track evaluative truth, then they are either not related to evaluative truth at all, or the correlation between evaluative judgments shaped by these processes and evaluative truth is negative. Both views, Street claims, are not viable due to their implausible skeptical implications.

Are *naturalist* realists about value in a better position to explain the relation between the selective pressures that shaped our evaluative attitudes and evaluative truth (which, on their view, consists in correspondence with natural facts)? In other words, are there more plausible ways to explain how evolutionary processes shaped our evaluative attitudes such as to track natural moral facts, which are nevertheless independent of these evaluative attitudes in the strong sense that which evaluative attitudes we do have has no effect whatsoever on the set of natural facts in which evaluative truth consists?[271] Street denies that such explanations exist, for the following reasons: In her view, the only way in which the naturalist value realist can proceed to find the correct "normative-natural identities"[272] (identify natural, mind-independent evaluative facts) is to start with our most stable evaluative judgments and try to make them as coherent as possible. However, the Darwinist will argue that evolutionary processes strongly influence even our most stable evaluative attitudes. Again, the realist has to give an account of how these processes relate to mind-independent evaluative facts, and the problems already noted reoccur. If there is *no* relation, it makes no sense to rely on evaluative attitudes in the search for natural evaluative facts, and no plausible alternative is conceivable. If there *is* a relation, the naturalist realist has to explain how tracking

[270] See ibid., p. 132.
[271] See ibid., pp. 136–137.
[272] Ibid., p. 139.

the natural evaluative facts made our evaluative attitudes evolutionarily advantageous. No such account seems to be forthcoming, and the adaptive-link account, which invokes the advantageousness of specific links between environment and evaluative attitudes, again appears to be the better theory. We make the evaluative judgments that we make (and not others) because they, or the evaluative premises they build on, constituted evolutionarily advantageous response tendencies tailored to specific circumstances. No mind-independent natural moral facts are required.

How does the antirealist (or antiobjectivist, in Kahane's terminology) avoid this Darwinian Dilemma while still being able to account for evaluative error? She argues that the evolutionary processes that shaped our evaluative attitudes and evaluative truth are related and that this relation is what the best available science describes (possibly, something like the adaptive-link account). The realist and the antirealist differ in their understanding of what Street calls the *direction of dependence* between the selective pressures that shaped evaluative attitudes and evaluative truth. The realist takes evaluative truth as prior and assumes that evolutionary processes shaped our evaluative attitudes so as to map these truths, while the antirealist takes evolutionary processes to be prior, shaping (among other influences) evaluative truth.[273] Note, finally, that Street understands her Darwinian Dilemma as a particularly acute version of a problem that any good scientific explanation of the shape of our evaluative judgments generates for realist theories of value. In any such case, a realist has to interpret the factors in this explanation either as distorting, which implies implausible skepticism if the scientific account purports to describe decisive influences on our judgment, or else as related to evaluative truth in a way that made our evaluative attitudes truth-tracking. Street adds that antirealism about value can accommodate notions of evaluative error even if "the standards for such errors are ultimately "set" by our own evaluative attitudes."[274]

In sum, Street rejects objectivism because it either has highly implausible consequences or else has to posit obscure theories about how tracking mind-independent evaluative truth is evolutionarily advantageous, while better nonobjectivist accounts are available. Regarding Kahane's exposition, one might concede that "evolution is not a truth-tracking process with respect to evaluative truth" in Street's account. This is the case, however, because there is no evaluative truth in the mind-independent sense of Kahane's epistemic premise.

[273] See ibid., p. 154.
[274] Ibid., p. 156.

The evolutionary debunking in the work of Greene and Street serves distinct undertakings. Street's argument 'debunks' the view that there is a mind-independent evaluative truth by appeal to the explanatory power of evolutionary accounts of the development of evaluative attitudes and the implausible *global* consequences of adopting the mind-independent view given her assumptions about the pervasiveness of evolutionary influence. The global reach Kahane presents as a problem for EDAs is in fact precisely the consequence of an objectivist view Street relies on to expose the implausibility of this doctrine. Greene's debunking, in contrast, is based on impressions of moral irrelevance triggered by confrontation with specific factors that affect *some* moral judgments due to evolutionary processes. These *targeted* debunking arguments draw their appeal from immediate impressions of the sort "*That factor* is not something moral judgment should be sensitive to!" Such relevance judgments can be (and are in fact) made without explicit accounts of what factors or processes should determine moral judgments, i.e., without reference to mind-independent moral truth.

Greene and Street make different assumptions about the pervasiveness of evolutionary influences. Street holds that all moral judgments are to some extent owed to evolutionary processes, while Greene believes that only *some* moral judgments, namely those characteristic of deontological ethics, are vulnerable to the evolutionary critique.

## 2.7 Kumar and Campbell on the Normative Significance of Moral Psychology

Victor Kumar and Richmond Campbell develop a model of how findings in experimental moral psychology can be normatively significant. They claim that Greene's findings cannot undermine deontology in general. However, the findings can be normatively significant if combined with plausible normative assumptions, which Greene fails to provide.[275] Consider Greene's argument from morally irrelevant factors: Empirically, the difference between personal and impersonal harm appears to be causing the difference between judgments about the footbridge and trolley cases. Judging pushing impermissible in the footbridge scenario counts as characteristically deontological, while the permissibility judgment in the switch case is characteristically consequentialist. In Kumar and Campbell's interpretation, Greene's argument reads as follows:

[275] See Kumar & Campbell (2012), pp. 311–312.

| | |
|---|---|
| Descriptive premise | The deontological judgment about footbridge is a response to personal harm. |
| Normative premise | Personal harm is a morally irrelevant factor. |
| Conclusion | The deontological judgment about footbridge is unwarranted.[276] |

This is an argument against deontology because the majority judgment in footbridge is thought to support deontology; it is not consistent with consequentialism. In conjunction with the second premise, the empirical evidence shows that the majority judgment is not well founded, and can thus no longer serve as support for deontology.

Next, Kumar and Campbell identify a moral-epistemological assumption that is not made explicit, but in their view nevertheless crucial to Greene's argument: "[T]he principal evidence for moral theories is our first-order intuitions about concrete cases. One moral theory is more justified than another principally insofar as it better explains and systematizes our first-order intuitions."[277] Without this assumption, they argue, Greene could not claim that his argument undermines support for deontology by undermining the trustworthiness of an important intuition with which deontology, but not consequentialism, is compatible. If it were not for this assumption, deontologists might respond that they do not care about correspondence with intuitions, since deontological principles are otherwise justified.[278]

Section 3 of Kumar and Campbell's paper is a reconstruction of Berker's argument against Greene. Contrary to Berker's own assessment,[279] they take his most promising objection to Greene to be that consequentialist judgments might be just as susceptible to psychological debunking as are deontological judgments, for instance, if intuitions at the base of consequentialist judgments are themselves affected by morally irrelevant factors (or unaffected by relevant factors). At this stage, Berker claims, a stalemate between opposing intuitions about what factors are morally relevant occurs and blocks any advance in the

[276] Ibid., p. 313.

[277] Ibid. Greene (and Singer) would probably say that deontologists measure the quality of their theory by correspondence with intuitions, but that consequentialists do not.

[278] Kumar and Campbell concede that what Greene argues for explicitly is the psychological claim that deontological thinking and consequentialist thinking are grounded in two separate psychological systems, and in particular, that deontological positions are based on emotional intuitions. While this is not equivalent to the philosophical claim that intuitions are the principal justification for moral theories, they claim that such an assumption is required in order to prevent the deontologist from responding that he does not care about how his position came about, because it is otherwise justified. See ibid., p. 327, note 7.

[279] See Berker (2009), p. 325: Berker's "most pressing worry" is that Greene's neuroscientific results are doing no work in his normative argument.

debate between deontology and consequentialism. Kumar and Campbell argue that this objection to Greene fails for two separate reasons:

Firstly, the suspicion that consequentialist intuitions respond to morally irrelevant factors is, at this point, mere speculation without empirical support. More importantly, Berker's objection is question begging. The moral relevance of the factors to which consequentialist judgments do *not* respond (e.g., the separateness of persons) *is* highly controversial between deontologists and consequentialists. The claim that "personal harm"[280] is morally irrelevant, in contrast, is not (at least according to Greene and Berker), which is why it is a potential basis for progress in the debate. Berker ignores this crucial difference, yet it is the reason why appeals to factors only deontologists consider morally relevant do not achieve much in their arguments with consequentialists.[281]

In order to formulate their own, "decisive" objection to Greene, Kumar and Campbell introduce "moral consistency reasoning"[282]. Consistency reasoning requires that we treat alike cases that are alike in all morally relevant features.[283] While attempts to reach reflective equilibrium focus on inconsistencies between moral judgments about specific cases on the one hand and moral principles on the other, moral consistency reasoning addresses inconsistencies between specific case judgments. The impression of inconsistency can arise when we perceive no morally relevant difference between two cases that elicit different judgments. There are two ways to resolve it: Either drop the less tenable judgment, or identify a morally relevant difference.[284] Judith Jarvis Thomson's famous-violinist argument for the permissibility of abortion serves as an example: The argument builds on the observation that there seems to be no morally relevant difference between unplugging and thereby killing a famous violinist hooked up to you circulatory system against your will, and aborting a fetus conceived in an act of rape. If no morally relevant difference between these cases emerges, disconnecting the violinist and aborting the fetus should be judged equally. Thomson uses this argument in defense of abortion, since the majority of people judge that it is permissible to disconnect the violinist. In order to defend abortion, one has to maintain that the judgment about the violinist case is more reliable, thus, we should align the evaluation of abortion in cases of rape with it.

---

280 I believe that Kumar and Campbell's identification of the morally irrelevant factor is mistaken. It should be 'personalness of harm'.

281 See Kumar & Campbell (2012), pp. 314–315.

282 Ibid., p. 315. See also chapter 5.5.3.

283 See ibid.

284 See ibid., p. 316.

According to Kumar and Campbell, both Greene and Berker agree that the normative premise "personal harm is a morally irrelevant factor" is uncontroversial. In fact, they argue, close consideration of this premise shows that Greene is in a dilemma: In the stated form, the premise is false, and modifications that remedy this deficit no longer lend support to Greene's conclusion (that the deontological judgment about footbridge is unwarranted). The normative premise is false because personal harm *is* morally relevant, and a good reason for moral disapprobation. I find this criticism far-fetched, since both Greene and Berker have clearly argued that the *personalness* of harming is morally irrelevant, and not that all *harm* is morally irrelevant if inflicted in a personal way. If personal harm were morally irrelevant, the footbridge case would not constitute a dilemma.[285] As a solution to this alleged problem, Kumar and Campbell suggest a similar formulation: "Judging that it is impermissible to push the large person in *footbridge* because the harm is *personal rather than impersonal* is incorrect."[286] However, this strategy likewise encounters difficulties: "It is equally true that judging that it is permissible to flip the switch in *bystander* because the harm is *impersonal rather than personal* is incorrect."[287] The correct thought behind the normative premise, they take it, is that the *difference* between personal and impersonal harm is not a morally relevant difference, and hence does not justify differential judgment. Given that no other morally relevant difference is discovered, it follows that *either* the judgment in footbridge *or* the judgment about bystander requires revision.[288] It cannot be inferred from the premises *which* of these judgments should be revised. Therefore, Greene's findings do not undermine just the judgment of the majority in footbridge and thereby reduce support for deontology. They do, however, (in combination with the normative premise that the difference between personal and impersonal harm is not morally relevant) cast doubt on moral principles that match the majority judgments about footbridge and bystander (e.g., the DDE).[289]

Greene could substantiate his attack on deontology by providing reasons why the bystander judgment, rather than the footbridge judgment, should remain unchanged. According to Kumar and Campbell, mere appeal to the 'obvious' adequacy of a permissibility judgment in bystander will not suffice, because that judgment is not, in fact, obvious to all in all circumstances. For instance, the permissibility judgment in bystander is susceptible to an

[285] See Berker (2009), p. 321 "P2. The factors that make a dilemma personal rather than impersonal are morally irrelevant" and Greene (2010), p. 12: "In P2, I would simply say that "personalness" in the above sense is morally irrelevant."
[286] Kumar & Campbell (2012), p. 317.
[287] Ibid.
[288] Kumar and Campbell do not mention the option of revising both judgments.
[289] See ibid., pp. 317–318.

order effect: Subjects that judge footbridge before bystander are more likely to find pulling the switch impermissible, while order does not affect the impermissibility of pushing the large man in footbridge.[290] The conclusion that both judgments are not jointly warranted concerns not the *truth* of these judgments, but their *epistemic status*: Linking differences in judgment to morally irrelevant differences in features of the cases implies that judging them differently is unwarranted (given the information). There could be undetected morally relevant differences that affect the moral status of these cases. However, if we assume that moral theories are justified primarily by their correspondence with moral intuitions (as Kumar and Campbell believe Greene's argument requires), we can only know about the morally relevant through our intuitions. Given that assumption, they believe that our best option to find out what's 'really' relevant is to start from our intuitions about specific cases and then check for morally irrelevant influences on these intuitions.[291] In sum, Greene is pursuing the most promising approach, but misstates the conclusions it currently warrants.

To complete their critique of Greene's work, Kumar and Campbell set out to defuse his "other debunking arguments" against deontology. Greene's dual-process theory claims that deontological judgments in moral dilemmas are the output of system 1, while consequentialist judgments stem from system 2. One way in which Greene argues that we should discard the impermissibility judgment in footbridge in order to remedy the inconsistency is by pointing to the unreliability of system-1-generated moral judgments. To this end, one might appeal to several characteristics of system 1: It is emotional, rather than rational. Moreover, as an evolutionary adaptation, system 1 produces output that was fitness enhancing on average, but not necessarily adequate moral judgments. Finally, system 1 processes are often described as heuristics. Heuristics approximate optimal solutions, but their efficiency depends on the relation between the heuristic process and environmental conditions. If any of these characteristics or their conjunction imply that the footbridge judgment is unreliable, this support of deontology is undermined.[292]

> Greene nowhere presents a compelling case that [the facts that system 1 involves emotional processing, was shaped by natural selection, employs simple and inflexible heuristics] reflect normatively appropriate criteria of evaluation […]. He does not tell us why emotional processing is worse than reasoned processing, or why

290 See ibid., p. 318.
291 See ibid., pp. 318–319.
292 See ibid., p. 319.

> fitness is unlikely to be correlated with moral truth, or why the simplicity and inflexibility of a rule impugns its content when applied to trolley cases.[293]

Kumar and Campbell discuss the significance of these characteristics by considering the heuristic character of system 1 processes with reference to Greene's 'camera analogy' for moral judgment: Automatic systems work well (take good pictures) in familiar circumstances (those the automatic system was designed for), but are unlikely to generate desirable output in unfamiliar situations. Kumar and Campbell argue that this does not support Greene's attempt to undermine deontology by reference to judgments in trolley cases. To achieve that, the footbridge scenario would have to be a fundamentally new moral problem, in which automatic emotional responses are not to be trusted. However, Greene does not claim that (they believe such a claim would be rather hard to defend)[294]. Instead, he mentions climate change and global poverty as instances of fundamentally new moral problems; these are indeed better candidates for situations that have no counterparts in the environment to which our automatic propensities to judge adapted. Nevertheless, because Greene does not argue that the footbridge dilemma is a fundamentally new problem, he cannot substantiate his claim that we should not trust system-1-generated judgments in that case and therefore adjust the judgment in footbridge to match that in the bystander dilemma. Thus, there is currently no reason to believe that automatic judgments are unreliable in trolley cases (although there might be such a reason, unbeknownst to us).[295]

What general lessons do Kumar and Campbell draw? Firstly, if Greene's empirical results are correct and the normative premise 'the personalness of harm is morally irrelevant' is uncontroversial, the classic *pair* of moral intuitions on footbridge and bystander is unwarranted, diminishing support for moral theories insofar as they incorporate the DDE or other principles that match these intuitions. Reality, however, is more complex: Factors beyond personalness or impersonalness affect moral judgment. One difference between bystander and footbridge that affects judgment is that harm is *intended* in footbridge, while it is merely *foreseen*, but not intended, in bystander. More precisely, subjects care about intentions particularly in cases of personal harm, while intentionality is not as important in

[293] Ibid., p. 320.

[294] Greene believes that the footbridge situation is fundamentally new insofar as situations in which several individuals can be saved by sacrificing one were not common in the EEA. Personal communication, April 2012.

[295] See ibid., p. 321.

the evaluation of impersonal harms.[296] This means that the DDE captures part of the determinants of moral judgments. In any case, even if Greene's empirical results do not amount to a complete explanation of the causes of moral judgment in trolley cases, the argument from morally irrelevant differences is valid and shows that empirical results can be normatively significant.[297] Kumar and Campbell also formulate methodological results: Moral philosophy uses consistency reasoning mainly to criticize widespread judgments on a particular issue A by referring to a more stable, yet opposing, widespread judgment on some other issue B. It is then argued that there are no morally relevant differences between A and B, and that consistency should be established by revising the judgment on A.[298] Empirical research can identify those differences between both cases that cause the difference in responses—the "psychologically efficacious difference"[299]. It does so by comparing responses to so-called 'minimal pairs' which (presumably) differ in only one respect. Neither the conclusion that the judgments are not jointly warranted, nor the suggestion which judgment to revise are final, since hitherto unknown morally relevant differences might be discovered in the future. Singer's shallow-pond problem illustrates this pattern: Many deny that help for the faraway needy is obligatory (while nevertheless praiseworthy) (A). On the other hand, it seems quite clear that, under normal conditions, everyone is obliged to rescue a child drowning in a nearby shallow pond (B). Is there a morally relevant difference between the child in a nearby pond and a starving child far away? Singer denies this, and argues that we should therefore revise (A). Experimental evidence can support his cause by identifying the psychologically efficacious difference between both cases, if that difference is clearly morally irrelevant. In the case of the starving and the drowning child, physical distance seems to have the largest effect on judgment, and Kumar and Campbell argue that differences in distance are not per se morally relevant.[300] They also agree that the judgment about starving children far away, not drowning ones nearby, should be revised (sadly, without giving further reasons for this choice). What about Thomson's argument in support of the permissibility of abortion? She argues that the judgment that "abortion of a fetus conceived in an act of rape is wrong" (A) should be revised because several facts coincide: The assessment is inconsistent with the judgment that "disconnecting oneself from a famous violinist that was hooked up to one's vital systems against one's will is permissible" (B), judgment (B) is more stable, and there is no morally relevant difference between the cases.

296 See ibid.
297 See ibid., p. 322.
298 See ibid.
299 Ibid.
300 See ibid., p. 323.

In contrast to the shallow-pond example, psychological research on this pair of judgments does identify a psychologically efficacious difference whose moral irrelevance is *not* uncontroversial: Studies indicate that the presence of a familial relation between the two connected individuals affects judgments. When the vignette states that the violinist is a half-brother, more subjects judge that disconnecting him is morally wrong.[301] Whether familial relations are morally irrelevant is controversial. Thus, these empirical results do not undermine unequal treatment of the two original cases. Rather, they suggest that people are responding to a potentially relevant difference. Kumar and Campbell conclude that research findings about intuitive judgments can advance ethical debates in which the consistency of these judgments plays a crucial role, if these findings can draw on uncontroversial assessments of the moral relevance of psychologically efficacious differences.[302]

## 2.8 Summary, and the Necessity of a Theory of Moral Judgment

Even though these findings on action and judgment probably involve various psychological mechanisms because the situations and tasks considered are quite diverse, they have a common message: Moral behavior and judgment appear fickle. Day-to-day decisions on how to act seem to be determined by irrelevant factors, and prone to ignore morally relevant features of the iudicandum. The situationist picture of behavior is unflattering.[303] With respect to judgment, intuitive emotional responses to moral issues might not provide the degree of reliability and consistency we would like to characterize our moral evaluations, even if some do relate to morally relevant factors. We conceive of moral judgments as differing in adequacy, where inadequacy can spring from influence of irrelevant factors and faulty balancing of relevant factors, including failure to consider them. Common sense attributes some importance to reason and argument in arriving at adequate verdicts, but apparently, not every moral judgment incorporates such rational influences.

Let us recapitulate the positions taken by the authors discussed regarding the significance of moral psychology. Singer sees his general distrust in intuitions strengthened by recent findings. Intuitions owed to evolutionary or cultural history are not to be trusted, and moral psychology is apt to identify them. Thus, evolutionary moral psychology promises a clearing up of the landscape of moral judgments. It might also identify those 'rational' psychological processes that supposedly produce adequate judgments. Berker, in contrast, targets neuroscience and claims that it has no normative significance. According to him, a normative

301 See ibid., p. 324.
302 See ibid., p. 325.
303 See Doris & Stich (2005), pp. 139–140.

intuition that is not a result of moral psychology does all the work in Greene's argument from morally irrelevant factors. At best, he believes, neuroscience can identify brain areas active both in cases of clearly fallacious reasoning and in moral reasoning and thereby prompt us to scrutinize the reliability of these judgments. In response, Greene points out that moral psychology provides new premises for arguments from irrelevant factors: Without empirical science, we would not know about the influence of factors such as personalness. Thus, moral psychology can supposedly help us find out which intuitions are reliable. Moreover, it can also point to evolutionary influences in or the emotional and heuristic character of moral intuitions, all of which render them unreliable at least in 'fundamentally new' situations. In those situations, we should rely on cognitive, as opposed to emotional, psychological mechanisms. Kahane worries that evolutionary debunking arguments as proposed by Greene and Singer have global reach and in fact affect all kinds of moral or evaluative judgments, since he suspects that all of them are affected by evolution. Moreover, he believes that these debunking arguments presuppose the truth of objectivism. Street, in contrast, believes that widespread evolutionary influence in our moral judgments renders objectivism implausible, since such a position would have to claim either that most of our moral judgments are false, or that evolution tracks moral truth. Evolutionary moral psychology could thus help decide metaethical questions. According to Kumar and Campbell, empirical research on moral judgment can become normatively significant if it identifies uncontroversially irrelevant psychologically efficacious differences between cases. Opposing judgments regarding such cases are then jointly unwarranted. These findings require further normative argument regarding which of the conflicting judgments to discard. However, experimental results do not warrant the judgment that some difference is not morally relevant, nor do they provide normative arguments for discarding one of the judgments.

At this point, I want to introduce a distinction between two kinds of judgments, namely first-order judgments that concern the moral status of actions or situations (good, bad, permissible, etc.) and second-order judgments about the moral *relevance* of factors that influence first-order judgments. I believe that we need a descriptive account of both first- and second-order judgments in order to assess properly the significance of psychological research about moral judgment. I propose such an account in the following chapters.

Why do we need such an account? It seems that the potential for irritation of both the results regarding behavior and those regarding judgment can be traced back to their relation with notions of moral relevance: The experiments show that moral judgments and behavior with a moral dimension are prompted by apparently morally irrelevant factors, or fail to

respond to apparently relevant factors (i.e., over- or undersensitivity). Not only are more 'rational' faculties sometimes ineffective in moral *judgment*, but the motivation to obey established moral rules of conduct can be silenced by insignificant happenstances and distorted to an alarming degree.[304] When neuroscientific and evolutionary explanations supplement such findings, criticism does not merely point to the inadequacy of the incriminated judgment, but includes these varyingly proximate explanations of judgment and behavior, as in "[moral] intuitions are prompted by features of our mind/brain that […] cannot be taken to be reliable guides to moral reality"[305]. However, commentators often offer only negative statements of the type "whatever proper behavior/moral judgment is, it is not *that*", combined with rather blurry proposals on how to remedy the deficits, rather than positive definitions of moral relevance.

The picture of morality which emerges in the following chapters will not only provide a better understanding of moral relevance, but will also show to what extent morality is shaped by intuitive, emotional, heuristic, and evolved psychological processes. On that basis, we will be able to assess the different suggestions regarding the normative and metaethical relevance of scientific approaches to morality addressed in the preceding chapters. The descriptive theory of first- and second-order judgments proposed latches onto all the positions regarding the normative significance of moral psychology so far presented. Insofar as arguments by Greene and Singer tie the trustworthiness of moral judgments to the degree to which they are shaped by evolutionary influences, they are subject to Kahane's worries about an unanticipated expansion of the scope of these arguments in case evolutionary influence on moral judgment is more widespread than Greene and Singer suspect. The actual reach of evolutionary explanations is crucial also for the assessment of Street's argument against objectivism. Therefore, the following chapters will point out evidence for such evolutionary components in moral judgment. This will also help gauge Berker's suggestion that consequentialist judgments are owed to evolutionary history as well. While Singer did not specify what more rational, reliable processes of judgment look like, a better understanding of the psychology of moral judgment might provide us with the tools to answer this question. I sympathize with Berker's emphasis on the role impressions of moral relevance have in arguments for alleged normative consequences of empirical findings. If moral relevance is such an important lever in normative argument, it is worthwhile to investigate

304 See ibid., p. 119.
305 Levy (2007), p. 288.

it more closely from a psychological perspective. Could it be that not all such impressions are equally reliable?

Greene's argument for the significance of empirical science seems convincing: Science provides new premises that, if combined with normative intuitions about moral relevance, yield new normative conclusions. Therefore, a better understanding of the determinants of first-order moral judgment *and* of notions of moral relevance can give us a better grasp of conceivable empirical premises. I will also discuss in detail the extent to which *emotions* shape both first- and second-order judgments, a characteristic Greene takes to undermine the trustworthiness of intuitions in fundamentally new situations. Because Greene thinks similarly about the potentially heuristic character of moral intuitions, I will consider to what extent moral intuitions are heuristics in part III. In the course of describing the mechanisms behind first- and second-order judgment, we might also attain more clarity as to the different levels of explanation at which impressions of irrelevance may arise. Moreover, a comprehensive descriptive account of moral judgment and moral relevance hopefully will provide the tools required to assess the second challenge posed by Kahane, namely whether debunking arguments require an objectivist or realist outlook on morality. Remember that according to Street's dilemma for objectivists, objectivism becomes less plausible if our moral judgments are saturated with the influence of evolutionary processes. Maybe we find that the reasons why we consider some judgment-producing processes to be 'off track' can be explained (and justified?) in a subjectivist spirit once we have in hand a sufficiently precise understanding of moral psychology. Finally, Kumar and Campbell's notion of moral consistency reasoning also relies heavily on judgments of moral relevance, or rather, irrelevance. Might moral psychology enable us to go beyond the assessment that a set of moral judgments is jointly unwarranted, and help us to figure out which judgment we should adjust? Assessing these issues requires, in my view, a deeper understanding of moral psychology.

Presumably, no normative moral theory is completely detached from intuition. To some extent, every theory trusts that the value of its central aims is evident without further argument; otherwise, attempts at justification would continue infinitely. Empirical evidence, however, indicates that such intuitions might be systematically misguided.[306] Is moral philosophy in trouble? Answering this question requires dealing with several subordinate issues. How exactly does moral judgment work? When is it susceptible to unwanted influences, and why? It seems evident that some factors should play no role in moral judgment,

[306] See ibid., p. 281.

while the relevance of others is a matter of debate. However, *how* do these differences in the consensuality of relevance evaluations arise? Why is the moral irrelevance of certain factors evident? Whence do impressions of relevance derive their legitimacy? Are there alternatives to founding ethics on intuitions? The following four chapters present further psychological findings and perspectives on moral judgment and action. Although the emerging picture is still quite sketchy, it allows for some educated guesses regarding answers to the questions just posed. The investigation into the nature of first- and second-order judgments commences with a distinction typical of the cognitive-developmental tradition in psychology: the discrimination between the *moral* and the *conventional* domain.[307] Many individuals feel they can discern moral rules from social conventions, and thus moral transgressions from violations of convention. Based on this perception, it might be possible to list criteria that distinguish these two domains. More to the point, if moral rules differ from conventional rules with respect to the *issues* they deal with, criteria that separate these classes of issues might throw light on what it means for a characteristic of a situation or action to be morally relevant.

[307] See Prinz & Nichols (2010), p. 120.

# Part II

—

# Contemporary Moral Psychology and Moral Relevance

# 3 Psychological Conceptions of the Moral Domain

This chapter reviews psychological theories on the hallmarks of moral judgments in the hope of identifying the features of situations or actions which affect their moral evaluation (the morally efficacious), the mechanisms of first-order judgment, and the features people think should be considered (the morally relevant). Firstly, I portray the cognitive-developmental tradition that dominated moral psychology for around thirty years from the late 1950s onwards. After that, I discuss recent movements, including a focus on emotions as well as on similarities and differences across cultures.

In the wake of the shift away from behaviorism and its eponymous focus on observable stimuli and responses, towards investigations of cognitive processes ('thinking') and their development, psychologists construed moral judgment mostly in terms of conscious verbal reasoning.[308] Such reasoning dealt with balancing and justifying *moral* rules, in contrast with *conventional* rules. These types of rules and the corresponding judgments (whether some act constitutes compliance with or violation of a rule) were considered psychological natural kinds, characterized by clusters of content- and response characteristics. Recent research, however, puts a question mark over this distinction by showing that the clusters can come apart. Apparently, judgments fall into more than two categories with respect to both rule content and characteristic responses to rule violations. Moreover, the ascription of several criteria employed in the moral/conventional classification of rules and judgments differs across cultures and social strata. To the extent that a delineation of the moral domain depends on both subject matter *and* typical responses to violations of these rules, it should take such variations into account.

## 3.1 Harm, Rights, and Justice: The Morality-vs.-Convention Framework

The moral/conventional distinction is closely associated with research conducted by American psychologist Elliott Turiel and his colleagues in the 1970s and 80s.[309] Turiel was working within the cognitive-developmental tradition of moral psychology brought into being by Jean Piaget and amplified significantly by Lawrence Kohlberg. Kohlberg believed that the ability to distinguish alterable *conventional* rules from less flexible *moral* norms develops only at the postconventional level (chapter 1.1.2). This is where Turiel's work (the so-called

[308] See Turiel (2006a), p. 792.
[309] See for example Turiel (1982).

'social-interactionist perspective') plugs in[310]: He devised a research paradigm to show that children distinguish between moral and conventional rules much earlier, from around the age of five. Contrary to Kohlberg's claim that a *conventional* understanding of obligations *necessarily precede*s proper understanding of *moral* obligations, Turiel argued that capabilities to comprehend morality and convention develop in parallel.[311] To Turiel, the distinction between moral and conventional rules is psychologically real and innate.[312] However, recent investigations presented below suggest that this distinction is partly an artifact of experimental design.[313] Turiel assessed the presence of a moral/conventional distinction in a paradigm known as the 'moral/conventional task', which has been used with a large range of subject populations. Subjects are confronted with transgressions of 'prototypical' moral and conventional rules and questioned about the presence of criteria believed to distinguish both types of rules. Examples of prototypical *conventions* are gender-specific dress codes ("men do not wear dresses"), regulations of proper forms of address, or the rule that pupils should not speak in class until called on. Examples of prototypical *moral* rules are prohibitions of killing or hurting others, stealing, or breaking promises.[314] The following table lists the criteria thought to distinguish moral from conventional rules; characteristics considered typical of moral rules are just the negation of what is supposedly true of conventional rules.

310 See Haidt (2001), p. 816.
311 See Turiel (2006a), p. 827, also Haidt (2008), p. 67, Shweder et al. (1990), p. 2.
312 See Prinz (2014), p. 106.
313 See Kelly & Stich (2007).
314 See ibid., p. 352, also Turiel (1982), pp. 148–150.

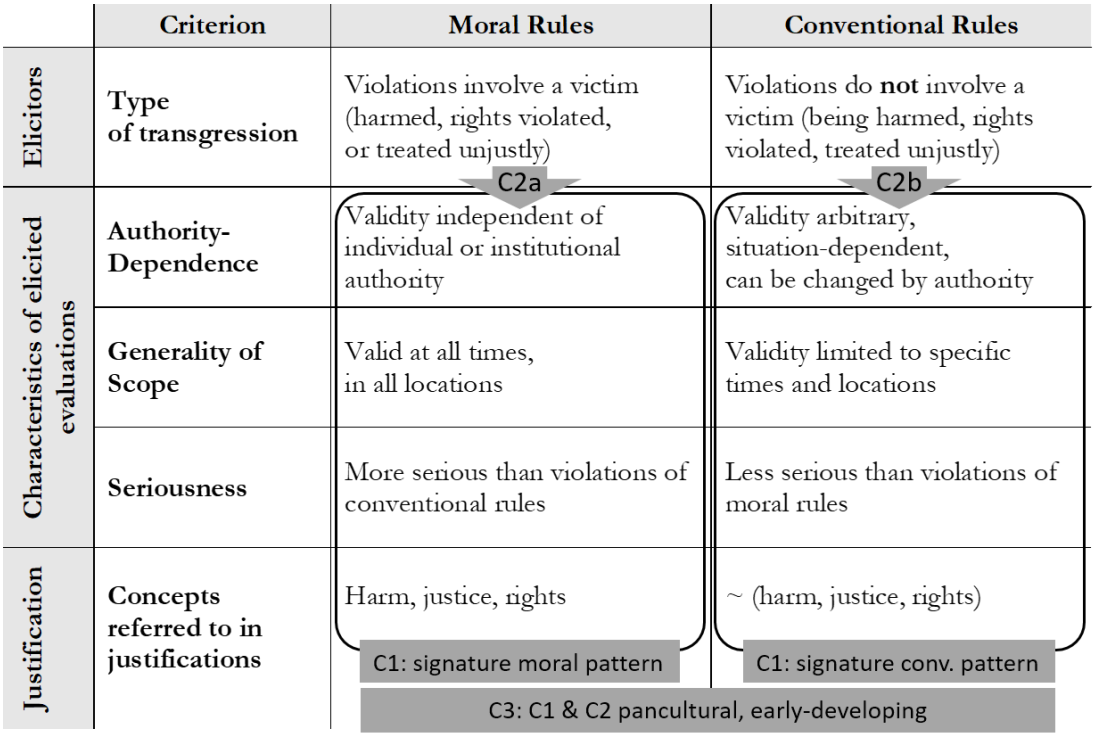


**Figure 3: Characteristics of Moral and Conventional Rules**

Based on Kelly & Stich (2007), pp. 352–355, illustration by BH

The characteristics of the transgression play a special role; violations of moral and conventional rules are supposed to elicit stable 'signature response patterns' respectively. Subjects regularly ascribe all three characteristics listed under 'moral rules' to rules transgressions of which involve harm, injustice, or a violation of rights (link C2a in the table).[315] They will say that the corresponding act (e.g., stealing, etc.) would be wrong even if it was allowed by authorities (authority independence), performed in the past, future, or an alien culture (universal scope), and that violations of the rule are a serious matter. These three criteria are reputed to form a "nomological cluster" (C1)[316]: Either all three are present, or none is. Thus, for all violations that do *not* involve a victim or any kind of harm, injustice, or violation of rights, the evaluation of the corresponding act is supposedly authority dependent as well as of limited validity and importance (link C2b). Whether or not a rule violation involves (or appears to involve) harm-/justice-/rights-violations is established by asking subjects how they justify the rule. If they refer to rights, justice or harm done, the rule counts

[315] I deliberately do not define the notions of harm, justice, and rights more precisely. The content of the following chapters gives reason to suspect that they are fuzzy concepts anyway.

[316] Kelly & Stich (2007), p. 355.

as moral rather than conventional.[317] In a sense, this is an attempt to define the morally efficacious by questioning subjects about the morally relevant. Indeed, a large number of studies found the predicted patterns of transgressions and responses across a wide variety of subject populations, with ages ranging from three and a half years to adults and including various nationalities and religious backgrounds. Even children with specific cognitive and developmental abnormalities, including autism, produced the signature transgression-response patterns.[318]

Do these investigations provide criteria for moral relevance? The hallmark of the moral/conventional paradigm is that *moral* transgressions typically involve a victim that has been harmed, treated unjustly, or whose rights have been disregarded.[319] It suggests itself that aspects of a situation that fit into these categories should form part of the morally relevant. A harm-focused notion of moral relevance is consistent with the perception that the presence or absence of personal force in the footbridge and switch dilemmas should not influence their assessment. Whether or not the victim is killed by use of personal force or by hitting a switch, the harm done is the same, thus, the moral evaluation should also be the same. The moral relevance of rights, on the other hand, fits the deontological impression that pushing the large stranger off the bridge is impermissible (for instance, if we think it would violate his right not to be used as a mere means to an end). On a different view, Greene's results cast doubt on the adequacy of the descriptive theory of the morally efficacious given expression in the moral/conventional distinction. After all, the moral/conventional framework is primarily an account not of how moral judgment *should* work, but of how it *does* in fact work. Thus if, as Greene suggests, moral judgments at least in some cases respond to factors (like personal force) which cannot (at least not without considerable effort) be understood in terms of either harm done, rights, or justice concerns violated, then the moral/conventional theory is descriptively inadequate in that respect.

---

317 Turiel actually distinguishes moral, conventional, and personal domain, but only the first two involve regulation. "Actions that do not entail inflicting harm or violating fairness or rights and that are not regulated formally or informally are consistent with the definition of the personal domain (these issues, in Western culture, include choices of friends, the content of personal correspondence, and recreational activities)." Turiel (2006a), p. 828. The consequences of actions in the personal domain primarily affect the actor. Accordingly, the conventional domain comprises actions that do not entail inflicting harm or violating fairness or rights, but *are* regulated formally or informally. Since this definition of the personal domain makes it less prone to evaluation by others, I do not consider it here.

318 See Kelly & Stich (2007), p. 355. Notably, the pattern is absent in psychopaths and children with psychopathic tendencies, who have difficulties to distinguish what others easily classify as moral or conventional rules. Rather, they tend to treat all violations in the manner normal individuals would treat moral rules.

319 Turiel defines the domain of morality as "prescriptive judgments of justice, rights, and welfare pertaining to how people ought to relate to each other" Turiel (1983), p. 3.

Greene's results alone do not show that the moral/conventional distinction is mistaken in claiming that all moral rules deal with matters of harm, rights, or justice. After all, the switch and footbridge dilemmas pose questions about harm and possibly rights, and they count as moral rather than conventional dilemmas. However, even if the danger of harm done, injustice, or infringements of rights is a necessary condition for an issue to be considered a *moral* issue, Greene's studies indicate that non-harm/-rights/-justice aspects affect judgment of acts which fulfill that necessary condition. Possibly, non-harm/-rights/-justice-related aspects are *not* systematically subordinate to potential harm, rights- or justice violations, i.e., issues could count as moral *even though* there is no victim in the harm/rights/justice sense. This is just what research by Jonathan Haidt and others suggests. The experimental paradigm employed in some of Haidt's studies is modeled on the moral/conventional task; the crucial difference being that transgressions were carefully designed *not* to involve any kind of harm, violations of rights, or justice concerns. The consensual-incest case illustrates this approach: Harms associated with incestuous relationships include an increased risk of handicapped offspring, disruption of the social order, or emotional harm done to the individuals involved. Haidt's scenario defuses these worries: Mark and Julie use two kinds of contraception (no risk related to offspring), the experience is enjoyable and makes their relationship more trustful (no emotional harm done), and they keep it a one-off and tell no one about it (no social disruption). Moreover, this consensual act does not seem to infringe upon anybody's rights; it is equally hard to discern any injustice. The vignettes nevertheless elicited a response pattern very much like the signature moral response pattern in subjects of low socioeconomic status both in Brazil and in the USA.[320] The respective acts were judged universally wrong and warranting interference. Apparently, rule violations that do *not* involve harm, injustice, or the violation of rights can provoke the moral response pattern (authority independence/universality/seriousness). Even so, the respondents typically tried to construct some kind of harm caused by the action when asked to justify their evaluations.[321] Actual moral judgments moreover appear not to depend on an (falsely) assumed presence of harm, since 'morally dumbfounded' subjects maintained their judgment even when they agreed that no harm occurred.

[320] See Haidt et al. (1993), p. 622.

[321] It is unclear whether subjects gave harm-related justifications merely to satisfy the interviewer or because they wanted to 'really' vindicate their verdict. The attempts grew increasingly helpless when subjects judged scenarios that excluded the most plausible sources of harm. One child justified his condemnation of cleaning the toilet with the national flag by stating that the flag might clog the drain. See ibid., p. 626.

Further studies challenge the immutability of the moral/conventional transgression-response patterns. Children in Arab villages in Israel showed moral response patterns in reaction to *all* rule transgressions presented even though several cases involved no harm, injustice or violation of rights (including calling a teacher by his first name).[322] In another study, North American children responded similarly to disgusting violations of etiquette rules (spitting into one's water before drinking it at a dinner party), while North American college students judged the wrongness of such transgressions to be authority independent and serious, but *not* universal in scope.[323] While the distinctions and probe questions used in these studies are not identical to those mentioned above, their results nevertheless indicate that the complete moral or conventional response patterns do not emerge as predicted, and that they are not the only possible patterns. Haidt et al.'s criteria to distinguish moralizing from conventional responses are quite similar to the moral and conventional signature response patterns as formulated by Kelly and Stich: Haidt et al. asked subjects about the necessity to interfere with the respective action and about the universal validity of the respective rule.[324] Accordingly, there were four response patterns resulting from combination of the answers. *Fully moralized*: endorsing interference and universalizing; *fully permissive*: opposing both interference and universalizing; in addition there were *enforceable-conventional*, i.e., endorsing interference while opposing universalizing (frequent response to a story in which a boy wears regular clothes in a school where pupils are supposed to wear a uniform) and *personal-moral*, i.e., opposing interference while endorsing universalizing.[325]

The results mentioned still allow that all violations of rules concerned with harm, justice, and rights evoke the signature moral response pattern, although the domain of morality extends beyond these matters; i.e., that violations involving harm etc. are *sufficient, but not necessary* to elicit the pattern. Examples of rule violations involving harm etc. whose evaluation is *not* taken to be authority independent, universal, and a matter of importance *all at the same time* would challenge this hypothesis. This possibility has not been investigated in detail; however, critics of the moral/conventional framework have presented initial evidence and arguments suggesting that harmful acts might not always elicit the signature moral pattern.

322 See Nisan (1987), quoted in Kelly & Stich (2007), p. 360. This example points to problems with the interpretation of the results: In other cultures, people might think that teachers have a *right* to be addressed by their last name, and conceive of this right as universal for lack of contact with cultures in which the rules are different.

323 See Nichols (2002), Kelly & Stich (2007), p. 361.

324 Haidt et al. (1993), 617: Interference: "Should [the actor] be stopped or punished in any way?" Universal validity: "Suppose you learn about two different foreign countries. In country A, people [do that act] very often, and in country B, they never [do that act]. Are both of these customs OK, or is one of them bad or wrong?"

325 See ibid., p. 622.

Firstly, many of the harm transgressions investigated thus far were of a narrow variety, typically such that they could happen in a schoolyard (pushing, hair-pulling, etc.). It might be premature to conclude that the same patterns will appear in response to all kinds of transgressions involving harm.[326] Secondly, there may be rules dealing with harm etc. which are *not* generalized to other times or locations (relativist attitudes), but nevertheless considered *moral*. Thirdly, rules prohibiting harmful treatment may not always be perceived as authority independent, as ongoing discussions about the legitimacy of torture to fight terrorism indicate.[327] Fourthly, variation could occur in the degree to which harmful transgressions are perceived as warranting interference. The evaluation of harmful acts is sometimes neither authority independent nor general in scope: Philosophers Stephen Stich and Daniel Kelly asked subjects whether it was OK for a teacher to spank a disruptive pupil. When the case vignette stated that the school's principal permitted spanking, acceptance rose to 48 percent, up from 8 percent in a vignette according to which spanking was explicitly prohibited. One might object that spanking is not a very serious form of harm. However, Kelly and Stich found the moral evaluation of more severe punishment to be similarly flexible: Another case asked for an evaluation of five lashes with a whip for a sailor who was drunk when he was supposed to be on watch. While only 6 percent of subjects thought this sort of punishment was acceptable on a modern American cargo ship, 52 percent thought it was acceptable on a cargo ship 300 years ago.[328]

Let us summarize. Since the moral/conventional framework provided the dominant delineation of the moral domain in psychology for the latter half of the twentieth century, analyzing it was a natural first step towards a psychologically informed notion of moral relevance and moral judgment. The framework advances hypotheses about the subject matter of morality and moral rules, as well as about the characteristics of responses to transgression of these rules. According to the moral/conventional tradition, morality deals with issues of harm, rights, and justice. Inflictions of harm, violations of rights, or injustices are supposedly considered wrong independently of whether they have been permitted by some authority, in all places at all times, and to be rather serious as compared to violations of conventional rules. The contentual thesis names concrete aspects of acts or situations that are both relevant to and efficacious in moral judgment. Empirical findings, however, sug-

326 Possibly, the signature moral response pattern occurs more reliably when harm is more serious.
327 See Kelly & Stich (2007), p. 362.
328 See ibid., pp. 362–365. The 'rule' in focus in this investigation is the rule not to harm others, not any rule regarding the particular duties of sailors aboard cargo ships.

gest that the matter is not as clear-cut. 1) Apparently, non-harm/-justice/-rights-related factors like the personalness of an act (switch vs. footbridge dilemma) affect the evaluation of harmful acts, regardless of whether people explicitly consider these factors relevant. 2) Harmless actions sometimes trigger the moral response pattern, including actions formerly considered transgressions of prototypically conventional rules. Haidt's moral dumbfounding results indicate that nonharm-related factors can be efficacious in moral judgment, even though the notions of moral relevance held by the subjects are limited to harm etc. The moralizing stance taken towards supposedly conventional transgressions like addressing a teacher by her first name, on the other hand, could indicate that subjects in non-Western cultures have a broader conception of morally relevant harm, or that their notion of moral relevance extends to matters beyond harm, rights, and justice. Possibly, these rule contents are not a *necessary* condition for moral responses. 3) Not all harmful acts elicit the entire moral response pattern: For instance, their wrongness is contingent on authority or time and location (Stich/Kelly experiments); thus, harm does not seem to be a *sufficient* condition for typically moral response patterns. Such responses might occur because these harmful acts do not count primarily as violations of rules prohibiting harm, but as being in accordance with rules that allow harming under specific circumstances. Moreover, 4) responses to rule violations have been observed that fit neither the signature moral nor the signature conventional response patterns (college students; *enforceable-conventional* and *personal-moral* responses). In the context of this investigation, results 1) and 2) are of particular interest: They suggest an influence of action characteristics unrelated to harm, justice and rights on moral judgment. Why were these factors neglected in the moral/conventional framework? Maybe people are unaware of them, do not consider them morally relevant, do not think that what actually affects judgment can be too distant from what they consider morally relevant, or assume that reference to those factors would not yield acceptable justifications.

## 3.2 Beyond Harm, Rights, and Justice: Insights from Cultural Psychology

The moral/conventional framework's claims regarding harm, justice, and violations of rights address both moral relevance and efficaciousness in moral judgment from a descriptive perspective.[329] In light of the results presented above, the moral/conventional framework appears inadequate as an account of what is efficacious in moral judgment, since

[329] Since the distinction between moral and conventional rules stems from accounts of moral development, it can also have a normative overtone, as it does, for instance, in the work of Kohlberg.

harmless actions elicit moral response patterns. Moreover, it might also be inadequate as a descriptive account of what people consider morally relevant. Some cultural psychologists, as opposed to exponents of the cognitive-developmental tradition, argue that the domain of morality extends beyond harm, rights, and justice in many societies.[330]

### *3.2.1 Shweder's Three Ethics*

In 1990, anthropologist Richard Shweder and his colleagues proposed a framework encompassing "three codes of moral thought and discourse, which cultures elaborate and rely on to different degrees"[331]. In what follows, I treat these 'codes' as families of issues considered morally relevant.[332] Harm, rights and justice are central only to the code of *autonomy* characteristic of Western, individualistic societies. Shweder et al. interviewed inhabitants of the Hindu temple town of Bhubaneswar, India, in order to gather judgments about thirty-nine potential breaches of codes of conduct. They then identified sixteen 'moral themes' subjects referred to in explaining their evaluations. Using statistical methods, these moral themes were condensed into three 'moral codes': the 'ethics of autonomy', 'ethics of community', and 'ethics of divinity'. These 'big three' coexisting discourses supposedly delineate the domain of morality. A single issue can pertain to one, two, or all of these discourses. Shweder et al. hold that an object of protection, particular regulative concepts, and a conception of the self characterize each discourse.[333] Moreover, each code relies on particular ontological assumptions:[334] Whereas the ethics of autonomy portray the world as populated by individuals, the ethics of community "sees the world […] as a collection of institutions, families, tribes, guilds or other groups."[335] Within the ethics of divinity, morality protects souls, and presupposes a divine entity and/or sacred order.

330 See Haidt et al. (1993), p. 613.
331 Ibid., see also Shweder (1990).
332 Shweder et al. (2003).
333 See ibid., p. 141.
334 See Haidt & Graham (2007), pp. 102–103.
335 Ibid., p. 102.

| | Ethics of Autonomy | Ethics of Community | Ethics of Divinity[336] |
|---|---|---|---|
| **Aim of protection** | Discretionary choice of individuals and the exercise of individual will in the pursuit of personal preferences | Integrity of stations or roles that constitute a 'community'; conceived of as entity with identity, history, reputation | Purity and integrity of the soul, the spirit, the spiritual aspects of the human agent and 'nature' |
| **Regulative concepts** | Harm, rights, and justice | Duty, hierarchy, interdependency, souls | Sacred order, natural order, tradition, sanctity, sin, pollution |
| **Conception of the self** | Self as individual preference structure | Self as office holder | Self as spiritual entity |
| **Means to achieve aim** | Increase choice and personal liberty. | Hierarchical superiors care for subordinates. They respond with loyalty and gratitude. | Avoid actions that separate the self from the divine unity. |

**Table 1: The 'Big Three' of Morality**

Based on Shweder et al. (2003), pp. 138–139

The discovery that morality extends beyond the issues emphasized in the cognitive-developmental tradition was sometimes accompanied by another observation: Shweder's interviewees appeared to have no appreciation of the 'conventional', but perceived their whole social order as a 'moral' order in which all rules were "universalizable and unalterable"[337]. Some authors have questioned the conclusion that the distinction between moral and conventional rules is not universal. Violations of a dress code and some nonreligious issues are treated more like conventional than moral transgressions by both Americans and Indians.[338] Shweder et al.'s inability to find typically conventional responses might have resulted from the fact that the breaches of codes of conduct they investigated did not include violations of a dress code.[339] Thus, the evidence indicates *not* that Indians lack the concept of conven-

336 "This moral code [divinity], with its emphasis on bodily practices, sounds strange and nonmoral to members of modern Western societies. Yet the ethics of divinity is highly elaborated in Hindu rules of purity and pollution […] and in the food, sex, and menstrual taboos of the Old Testament (cf. Leviticus 12-20)." Haidt et al. (1993), p. 614.

337 Ibid., see also Shweder et al. (1990), pp. 3–4.

338 See Turiel (2006b), p. 818.

339 See Shweder et al. (2003), pp. 131–135.

tional rules, but rather, that the realm of issues they meet with a moralizing stance (presupposing universal validity) is broader than in Western cultures like North America.[340] If the domain of morality comprises matters over and above harm, rights, and justice, signature moral response patterns triggered by acts that do *not* involve harm etc. in particular cultures (or socioeconomic strata, as the results of Haidt et al. indicate[341]) are no longer anomalous. Is this reclassification an improvement?

Notions of the prototypically moral often refer not only to *content* (harm, justice, etc.), but also to the ways in which those issues are dealt with, as exemplified in the concept of signature moral *response patterns*. Inferring the domain of morality from either set of characteristics *alone* will yield divergent classification of issues when matters become less 'prototypical'. A focus on *content* (harm etc., for example) loses plausibility when non-harm/-rights/-justice issues elicit the signature moral response pattern in a significant number of people. On the other hand, a focus solely on *response patterns* loses appeal if issues that fail to elicit the complete signature response pattern nevertheless often count as moral, or given that certain individuals, such as psychopaths, display the moral pattern in response to every rule violation they are questioned about. Apparently, stable intuitions about what belongs to the moral domain can always challenge the validity of theoretical concepts. Defining morality *merely* by saying 'whatever we consider moral, is moral', on the other hand, seems unsatisfactory. In what follows, I aim to strike a balance between intuitions and theory by exploring the ways in which 'what we consider to be part of the moral domain' is nonarbitrary.

The cognitive-developmental tradition contains a culturally biased notion of morality. Surely, notions of morality that characterize other cultures are equally important for a comprehensive descriptive account of morality.[342] Since what are, by Western standards, conventional rules elsewhere elicit 'moral responses', and notions of the morally relevant are often much more comprehensive, a culture-invariant meaning of 'morality' is difficult to

340 See also Haidt et al. (1993), 626.

341 Haidt's results regarding the influence of socioeconomic status on moral judgment might also indicate that the classical moral/conventional framework is an accurate notion of the moral domain or notions of the morally relevant of the (presumably liberal) Western academics who concocted it, but not an accurate picture of the moral domain for other social strata and cultures. See Doris & Stich (2005), p. 141.

342 This kind of approach, also present in Jonathan Haidt's writings, has been criticized on normative grounds: "Haidt appears to consider it an intellectual virtue to accept, uncritically, the moral categories of his subjects. But where is it written that everything that people do or decide in the name of 'morality' deserves to be considered part of its subject matter?" Harris (2010), p. 87. As regards descriptive accounts of morality, I believe that taking non-Western but elsewhere culturally established notions of morality seriously is indeed an intellectual virtue.

define in terms of either issues or response patterns.[343] The evolutionary origins and function of morality, and the cognitive processes executing moral judgment, could shed light on the heterogeneity of the phenomena subsumed under that heading. Differences between evaluations along dimensions like authority dependence, seriousness, or universality are matters of degree rather than binary, and may vary across cultures and individuals. Moreover, *many* issues considered relevant can be framed in the language of *several* areas of concern such as Shweder et al.'s 'three ethics'. There are no sharp dividing lines between these codes. While some see an act as harmful because they suspect it to be emotionally stressful, others might perceive a violation of the divine spirit present in each individual, or a display of disrespect towards particular roles in a hierarchy.[344]

What would it mean in terms of moral relevance and efficaciousness in moral judgment if the three-ethics framework were correct? Several scenarios are conceivable: Firstly and most obviously, there does not appear to be a notion of moral relevance that is equally valid and exhaustive of the moral domain for all cultures (possibly a significant 'ethics of autonomy' exists in most cultures, while the other codes are similarly important only in a smaller set of societies).[345] Secondly, regarding the *relation* of the morally relevant and the morally efficacious, Shweder's results are compatible with two constellations: 1) Both domains, the morally efficacious and what is considered morally relevant, differ across cultures, and different relations between the two sets of factors may exist. 2) The morally efficacious (actual determinants of moral judgment) is quite stable across cultures. Depending on the extent to which what people consider morally relevant determines what actually affects judgment, cultures which omit one, two or even, unlikely as it may seem, all codes in their notion of moral relevance might simply fail to appreciate some influences on their moral evaluations. Option 1 seems more likely than option 2 due to interdependencies between the notions of moral relevance prevalent in a culture and the actual determinants of moral evaluation. Moreover, the morally relevant is presumably often a subset of the morally efficacious.

Another important feature adds to the complexity a comprehensive account of moral evaluation should capture: Just as there are differences between what people in different cultures hold to be morally relevant, there are also differences between the notions of moral

[343] Nonetheless, they presumably consider some issues more serious than others (more or less authority-dependent or universal).

[344] Shweder et al. (2003), p. 142 point out how much notions of harm and rights have expanded in the United States. Harm includes such broad concepts as harassment, abuse, or exploitation; it encompasses phenomena as diverse as secondary cigarette smoke and stressful work environments. Children and even animals are thought to hold rights.

[345] See Graham et al. (2011), p. 380 for evidence that this is in fact the case.

relevance held by subgroups or individuals *within* larger cultural or regional groups. Think of religious and nonreligious individuals in Western cultures like Europe or the USA. Jonathan Haidt, the author of the studies on 'moral dumbfounding', has extended Shweder's three-ethics framework to capture notions of moral relevance in more detail and used this new framework to model different 'moral types' that can help characterize differences *across* cultures, but also map onto differences between political ideologies *within* cultures. The next section describes this so-called 'moral foundations theory'. Based on this theory, Chapter 4 illustrates in detail how emotions shape both the morally relevant and the morally efficacious.

### *3.2.2 Extending the Three Ethics: Moral Foundations*

Psychologists Craig Joseph and Jonathan Haidt surveyed the literature in order to identify the essences of moral relevance. Building on Shweder's work, they attempted to go beyond an analysis of discourse patterns and identify the origins of moral intuitions.[346] They included works covering aspects universal to moral systems all over the world, others dealing with how moralities differ between cultures, and research on possible precursors of morality in nonhuman primates.[347] As a result, they came up with five 'psychological foundations' that supposedly capture the fundamental categories of value around which the whole diversity of moral systems revolves. According to Haidt and Joseph, sensitivities for each of these areas, rooted in intuitive, affectively laden response patterns, evolved in answer to specific adaptive problems. It is useful to distinguish between the 'actual domain' and the 'proper domain' of evolved psychological mechanisms.[348] While 'proper domain' refers to the set of inputs the mechanism originally evolved to respond to (in other words, specific features of its EEA), its 'actual domain' denotes all inputs that *actually* trigger the mechanism. Especially in environments that differ substantially from the environment in which the respective mechanisms evolved, the actual domain may be quite unlike the proper domain. Consequently, it is a matter of empirical investigation whether responses to triggers in an environment other than the environment of evolutionary adaptedness, on average, bestow any benefit at all on the respective organism in terms of inclusive fitness. Accordingly, regular responses to input that is not part of the proper domain are by-products of the responses to triggers in the proper domain.

346 See Haidt & Graham (2007), p. 104.
347 See Haidt & Kesebir (2010), Haidt & Joseph (2004), and Haidt & Joseph (2007).
348 See Appiah (2008), p. 127, also Haidt & Joseph (2007), p. 381.

Haidt and Joseph assume that the similarity in moral (and other) valuations across cultures, as well as their cross-generational stability, are evidence of "some innate structure and content built into the mind."[349] They believe that much of this content is contained in domain-specific mental modules, because according to their research (e.g., on moral dumbfounding), moral judgment does not work as if it consisted in the domain-general application of principles to specific situations as imagined by Kohlberg.[350] While admitting for considerable modularity, Haidt and Joseph try to avoid two problems that beset massively modular conceptions of the mind: Moral valuation is more flexible (regarding the relation between triggers and responses) across cultures than other evaluative reactions typically associated with modules, such as fear of spiders or preferences for sweetness in food. This 'flexibility problem' could indicate that moral cognition *also* involves domain-general mechanisms that operate on broader classes of contents. Moreover, mental modules are often taken to be 'informationally encapsulated', i.e., not forwarding information to other modules. However, all kinds of additional information, manipulations of mood, etc. affect moral valuation; thus, it does not seem to fulfill this condition.[351] To accommodate these concerns, Joseph and Haidt assume 'moderately massive' modularity, a notion developed by anthropologist Dan Sperber: His so-called 'teeming' modules are to a large degree acquired during ontogenesis, highly variable, and frequently nested within each other. A smaller set of innate 'learning instincts' (modules) that regulate the acquisition of these domain-specific mechanisms governs the development of noninnate modules.[352]

> [F]or example, if there is an innate learning module for fairness, it generates a host of culture-specific unfairness-detection modules, such as a "cutting-in-line detector" in cultures where people queue up, but not in cultures where they do not; an "unequal division of food" detector in cultures where children expect to get exactly equal portions as their siblings, but not in cultures where portions are given out by age.[353]

---

349 Ibid., p. 378.

350 See ibid., p. 379.

351 See ibid., pp. 378–379. Note that the notion of EPMs introduced in chapter 1.2.3 does *not* require that mental modules be informationally encapsulated (see footnote 62), thus, the sensitivity of moral judgment to input from 'nonmoral' modules is, on my view, not at odds with a significant degree of modularity in moral cognition.

352 Pointing to the importance of innate learning dispositions, Tooby & Cosmides (2001), p. 14 distinguish between "actions produced to accomplish fitness-enhancing outcomes *in the external world*", "actions produced to cause fitness-enhancing changes to the *body*" and "actions produced to cause fitness-enhancing changes to the *mind/brain*" (my emphasis). Adaptations that produce the third kind of actions are 'developmental adaptations', designed to involve the individual in experiences which build and calibrate more mature mental adaptations. See ibid., p. 15.

353 Haidt & Joseph (2007), p. 379.

Each of the foundations of morality proposed by Haidt and Joseph could be a "Sperber-style learning module"[354], although they claim that what is crucial to moral foundations theory (MFT) is *some* sort of innate 'preparedness', not the specific hypothesis about modular architecture.[355] Table 2 lists each 'moral foundation', the adaptive challenge it evolved to respond to, its proper and actual domain, some associated emotions, and relevant 'virtues and vices'.

[354] Ibid., p. 381.
[355] See Haidt & Graham (2007), p. 106.

| | Adaptive Challenge | Proper Domain (adaptive triggers) | Actual Domain (set of all triggers) | Characteristic Emotions | Relevant Virtues [Vices] |
|---|---|---|---|---|---|
| **Care/ Harm** | Protect and care for young, vulnerable, or injured kin | Suffering, distress, or threat to own kin | Baby seals, cartoon characters | Compassion | Caring, kindness [cruelty] |
| **Fairness/ Cheating** | Reap benefits of dyadic cooperation with nonkin | Cheating, cooperation, deception | Marital fidelity, broken vending Machines | Anger, gratitude, guilt | Fairness, justice, honesty, trustworthiness [dishonesty] |
| **Loyalty/ Betrayal** | Reap benefits of group cooperation | Threat or challenge to group | Sports teams one roots for | Group pride, belongingness; rage at traitors | Loyalty, patriotism, self-sacrifice [treason, cowardice] |
| **Authority/ Subversion** | Negotiate hierarchy, defer selectively | Signs of dominance and submission | Bosses, respected professionals, subordinates | Respect, fear | Obedience, deference [disobedience, uppityness] |
| **Sanctity/ Degradation** | Avoid microbes and parasites | Waste products, diseased people | Taboo ideas (communism, racism) | Disgust | Temperance, chastity, piety, cleanliness [lust, intemperance] |

**Table 2: Five Foundations of Intuitive Morality**

Based on Haidt & Joseph (2007), p. 382, terminology from Haidt (2012)

First, there is the *care/harm* foundation, marked by a concern for the suffering of others. The second category, *fairness/cheating*, centers on issues relating to equality, fair treatment, and diverse notions of justice. Together, care/harm and fairness/cheating constitute a more

fine-grained reformulation of the ethics of autonomy in Shweder's framework, since they aim to protect the preferences and rights of individuals. Anthropologists assume that our ancestors lived in egalitarian bands as long as their lifestyle was nomadic, and that extensive hierarchies emerged only with the introduction of agriculture, food storage, and resulting possibility for large differences in wealth to accumulate.[356] The next two categories, *loyalty/betrayal* and *authority/subversion*, encompass matters of concern emphasized in ethics of community, which focus on the functioning of supraindividual social entities such as tribes, religious communities, nations, etc. More precisely, loyalty/betrayal concerns relate to "obligations of group membership, such as loyalty, self-sacrifice, and vigilance against betrayal"[357]. In contrast, authority/subversion deals with the moral aspects of hierarchy and social structure, such as living up to obligations associated with particular social roles, especially regarding protection, obedience, and respect. From the perspective of MFT, the dilemma in which subjects found themselves in Milgram's obedience experiments is thus a choice between two *moral* concerns: They could *either* avoid harm to the 'student' *or* obey authority. In contrast, Turiel, as a proponent of the traditional moral/conventional distinction, has interpreted the situation as a conflict between moral and *conventional* norms.[358] While Kohlberg's scoring manual implied that justifications with reference to authority (and tradition) are indicative of a conventional stage of moral competence and ideally overcome through increased role taking in the process of maturation, MFT specifies no such normative ordering of the different fundamental moral concerns.[359] The fifth foundation, *sanctity/degradation*, corresponds to Shweder's 'ethics of divinity' and subsumes norms about what body and soul should and should not come into contact with. Such considerations manifest, for instance, in norms regarding sexual behavior, health, and the control of urges and desires more generally.[360]

Recently, Haidt et al. have introduced 'liberty/oppression' as sixth foundation.[361] The corresponding adaptive challenge was to avoid bullying in small groups; all signs of at-

356 As a minimum, male hunters shared roughly equal prerogatives and obligations. See also Boehm (2012), pp. 80–81.
357 Haidt & Kesebir (2010), p. 822.
358 See Turiel (2006a), p. 834.
359 See Graham et al. (2011), p. 381.
360 See Haidt & Kesebir (2010), p. 822. Haidt has suggested that evolution works faster than expected, thus also the last 10000 years after the end of the Pleistocene could have shaped our inherited preferences. He also assumes that divinity concerns are much more recent than concerns about harm or reciprocity (http://symposia.templeton.org/darwin200/, talk 3, 12:50).
361 See Haidt (2012), pp. 170–176.

tempted domination are adaptive triggers. Indeed, anthropologist and primatologist Christopher Boehm considers bullying to be a more potent type of free riding than cheating.[362] A kind of righteous anger, also referred to as reactance, is the typical emotional response to these elicitors. It involves a motivation to unite with others in similar situations in order to take action against the oppressor. Boehm hypothesizes that such subordinate rebellions became much more frequent when humans took to hunting large game, because equal division of the irregular spoils was necessary to ensure constant sufficient nutrition to all group members. Fierce opposition to dominant behavior may have rendered the ability to constrain one's selfish impulses evolutionarily advantageous.[363] The class of actual triggers of reactance contains not only dominant individuals, but almost every entity perceived as illegitimately constraining liberty, such as government or its policies, or accumulations of wealth that are viewed as results of exploitation and abuse of power.[364] Assuming that the separate existence of this sixth foundation will receive more empirical support in the future, Haidt and his colleagues now suspect that concerns for *equality* are rooted in the care/harm and liberty/oppression foundations, while fairness/cheating concerns aim at *proportionality* rather than equality.[365] The dislike of oppression targets those who (illegitimately) amass resources, while the sensitivity for the well-being of others captured in the care/harm foundation enables us to care about those who (illegitimately) get less than an equal share. The fairness foundation is concerned not only with interactions among individuals, but also with what others contribute to tasks accomplished collectively.[366]

In Haidt and Joseph's framework, emotional intuitions, i.e., affective, quick, and largely automatic responses to situations are the evolved origins and mainsprings of morality. Sophisticated moral notions are the results of permanent cultural amplification and modification of the set of phenomena that trigger these affective systems. To render this account more tangible and persuasive, chapter 4 elaborates on how emotional responses function not only to establish fundamental moral concerns, but also to make people act in accordance with established moral codes. While MFT describes the moral domain more adequately than the moral/conventional tradition, it remains an intriguing question whether and if so, in which sense, the 'fundamental' concerns establish a special realm of issues.

362 See Boehm (2012), pp. 65–66.
363 See ibid., pp. 151–152.
364 See Haidt (2012), pp. 174–175.
365 See ibid., p. 180.
366 See ibid., p. 181.

### 3.2.3 Further Dimensions of Morality: Relational Models

Anthropologist Alan Fiske and psychologist Tage Rai propose a model of 'morality as relationship regulation' that integrates cultural, developmental, and social-psychological findings. Relational models theory (RMT) is an account of social relations first published by Fiske in 1991. The theory states that human beings in all cultures understand social relationships in terms of one of four relational models, and that these relational models determine which obligations and moral motives apply.[367] Importantly, the models attributed to specific relationships can differ across individuals and cultures. Some significant moral disagreements presumably result from the application of different relational models, implying different moral norms, to the same relation or seemingly similar relations, or from differences in the implementation of a given model, rather than from differences in knowledge or logical reasoning.[368] Rai and Fiske emphasize that in order to understand moral judgment, we have to dismiss the assumption that social context is irrelevant, or that social considerations introduce bias. Rather, context determines which norms apply and thereby affect moral judgment. Supposedly, it is a distinctive feature of their theory that *any* action can be morally acceptable, depending on which relational model and moral motive are active. At the same time, Rai and Fiske consider an action a genuine moral violation if there is implicit or explicit agreement within a group or culture regarding the adequate social-relational model in a given situation, and the act in question moreover contradicts the motives corresponding to that model.[369]

Rai and Fiske illustrate the effect of relational models with a striking example of a young woman in Syria who, after suffering abduction and rape, was upon her return stabbed and killed by her brother. Even though it is very hard to see from a Western perspective, Rai and Fiske argue that this brother was acting from moral motives.[370] "In its strongest form, a social-relational approach to moral psychology posits that the moral status of actions cannot be determined independent of the social-relational contexts in which they take place."[371] A weaker version of their thesis is to say that the moral status of iudicanda sometimes depends on social-relational context. Such an understanding of morality, they argue, departs from the post-enlightenment notion of morality as based on principles that are independent of social status or personal relationships. The cognitive-developmental tradition established

367 See Rai & Fiske (2011), p. 58.
368 See ibid. and Sunar (2009), p. 454.
369 See Rai & Fiske (2011), p. 68.
370 See ibid., p. 57.
371 Ibid.

by Piaget and Kohlberg is based on that understanding. Piaget thought of social constraints as hindering the development of autonomous morality; Kohlberg's highest levels are characterized by a universal understanding of morality, while social pressure constitutes a non-moral bias.[372] The distinction between moral and conventional norms in the social-interactionist tradition established by Turiel likewise views universality (authority independence) as a defining feature of moral norms, as opposed to conventional ones. These approaches in moral psychology did not take into consideration results of Milgram and other social psychologists that pointed to the potentially large effects of social relations on the permissibility of actions, helping behavior, or willingness to cooperate. Rai and Fiske consider these results and argue that "our sense of morality functions to facilitate the generation and maintenance of long-term social-cooperative relationships with others"[373]. In each relationship, there is some potential for exploitation. In order to maintain functioning relationships, people thus need suitable motives to regulate their own behavior, including motives to control the behavior of their counterpart. In this framework, Rai and Fiske attribute specific roles to emotions: Negative violations of the behavioral expectations that define a specific type of social relationship elicit aversive self- or other-directed emotions, depending on who transgressed (self-directed: guilt and shame; other-directed: disgust, envy, outrage). On the other hand, positive emotions like compassion, loyalty, and awe are closely tied to the specific obligations that characterize relationship types.[374] While Rai and Fiske state that aversive emotions lead to sanctions *after* the transgression occurred, it seems as if the positive emotions might both occur in response to norm conformity, as well as serve to motivate adherence to obligations in the first place. I suspect that the anticipation of aversive emotions directed both towards oneself and others can similarly prevent transgressions.

There are four basic mental models and corresponding motives for social relationships in relational models theory.[375] The communal-sharing (CS) model bases a relation on the perception that the other person shares an important feature with the self (the extent to which they have that feature does not matter, it is a binary categorization), such as belonging to the same family, team, nation, congregation, etc. The corresponding motives promote unity. Simplifying slightly, one might say that one treats those with which one is in a CS relation as if they were part of the self: If they are harmed or offended, the self feels harmed or offended; their needs have to be taken care of independent of considerations of merit or

372 See ibid., p. 58.
373 Ibid., p. 59.
374 See ibid.
375 See Sunar (2009), pp. 453–454.

reciprocity, property is shared among group members. If someone violates a norm, those who share a CS relation with the transgressor (sometimes even a whole group) feel responsible; the transgressor has to be cleansed of her guilt or expelled. The focus on unifying features shared between individuals is often accompanied by a negative connotation to deviant behavior, marked by disgust.[376] Rai and Fiske claim that in the honor-killing case, the differences in the moral assessment of the situation and its demands between Western and honor cultures result from different constructions of the CS model. While the relation between daughter and family is (also) a CS relation in both cultures, honor cultures understand rape as a violation and defilement of the group remedied only by killing or otherwise expelling the victim. In other, typically Western constructions of the CS relation, the focus is on the suffering of the victim and making her feel well again.[377]

When employing the authority-ranking (AR) model, people position other individuals on a specific dimension. Rights and duties vary as a function of status as they do, for instance, in the military or between parents and children, motives serve to maintain or establish hierarchy. In subordinates, hierarchy is maintained by obedience to the will of superiors and punishment of those who disobey orders, while superiors have a certain responsibility to guide and protect their subordinates. Since stable hierarchical systems regularly rest on some kind of reciprocity rather than one-sided coercion, hierarchies count as natural and legitimate in many cultures; relational models theory incorporates that observation. Hierarchies entail the expectation that those higher up are entitled to larger shares of valuable resources, but also to some extent accountable for the actions of their subordinates.[378]

In relationships structured according to the equality-matching (EM) model, people care about equality (motive) with regard to some specific unit, such as opportunity or satisfaction of needs, but also harm or damage done. The equality motive fits well with tit-for-tat strategies.[379] Market pricing (MP) is similar to EM insofar as a certain proportionality (motive) between interaction partners is called for. However, MP models employ ratios and rates to enable the balancing of distinct goods (exchange of money against goods is a prominent example), but also punishment that does not consist in eye-for-an-eye retribution. In EM relations, in contrast, people focus on the proportionality of one specific good.[380] MP relations and proportionality motives are in play when we weigh different goods to arrive at a decision. In general, Rai and Fiske state that the models become increasingly

376 See Rai & Fiske (2011), p. 62.
377 See ibid.
378 See ibid., p. 63.
379 See ibid., p. 64.
380 See ibid., p. 60.

complex from CS to AR, EM, and MP.[381] They therefore suspect them to emerge in this order in both ontogenetic and phylogenetic development.[382] It is important to note that in complex social relationships people often perceive different aspects of their relationship through different models. Moreover, each model can be 'constituted' in different ways, not all of which have to seem appropriate to all agents involved. Since the theory ties moral motives and obligations to social relations, it predicts that we are mostly indifferent towards those with which we share no relation at all.[383]

Rai and Fiske point out several aspects in which, in their view, the understanding of moral psychology as relationship regulation differs from existing moral-psychological theories: Unlike in the theories of Turiel or Hauser (chapter 5.3), violence does not have to result from nonmoral 'biases'. Unlike in Haidt's theory, there is no moral concern with purity that is independent of social relations.[384] Traditionally, violence seen from a moral point of view is often either an intentional violation of norms, a mistake, or justified by some greater good, but generally something to avoid. Rai and Fiske argue, however, that there are various forms of violence that are or were seen as not morally bad, even sometimes required, such as punishment for crime and disobedience, or violence afflicted in self-defense or on enemies more generally. Whether violence is permissible or even praiseworthy depends, according to Rai and Fiske, on the specific relational model within which it occurs. Within CS relations, violence against out-group members is generally more acceptable than violence against members of the in-group, and might even be praiseworthy if out-group members are perceived as a threat. Within AR relations, violence is more acceptable if exerted by superiors against subordinates than vice versa, and orders from superiors might even morally require subordinates to commit acts of violence.[385] EM relations require that violence exchanged between conflicting parties be sufficiently similar (lex-talionis style).

---

381 "[…] CS is homologous with nominal (categorical) measurement, wherein the organizing principle is group membership; formally, it consists of equivalence relations. AR maps onto ordinal measurement scaling, wherein the linear order of individuals is salient but differences cannot be quantified; mathematically, it is a linear ordering. EM corresponds to interval measurement, wherein differences can be added and subtracted to track imbalances; it has the structure of an ordered Abelian group. MP has the structure of a ratio scale with a defined zero point: It is an Archimedean ordered field […]." Ibid., p. 61.

382 See ibid., p. 68. Haidt (2001), p. 826 mentions evidence "that the four models emerge during development in an invariant sequence: communal sharing in infancy, authority ranking by age 3, equality matching around age 4, and market pricing during middle or late childhood."

383 See Rai & Fiske (2011), p. 64, where they refer to the notion of moral disengagement developed by A. Bandura.

384 See ibid., p. 65.

385 See ibid. Interestingly, Rai and Fiske refer to experiments in which participants morally judge the beating of a sailor by his captain on a seventeenth-century ship (similar to those described in chapter 3.1): In their view, that beating was judged more leniently because it was perceived as occurring in an AR relationship between captain and sailor. Presumably, one would have to refer to different constructions of AR relations to explain the judgment regarding the beating of a sailor on a contemporary ship.

Violence can also be weighed against other goods within MP relations. Even infants seem to be capable of 'praising' violent behavior if violence is exerted as a punishment for cheating: In an experiment, they preferred a puppet that punished a cheater to a puppet that helped the cheater.[386]

With respect to fairness, Rai and Fiske believe their model is better suited than an understanding of fairness as equality to capture what people actually perceive as fair in different social relationships. While equality might constitute fairness in EM relations, equal distribution of resources would probably *not* appear fair in AR relationships. The same is true for relationships governed by proportionality motives (MP), in which benefits and costs are supposed to be proportional to merit, capability, etc. Even within CS relations, concerns with equality can appear pedantic and opposed to the motive of giving everyone according to his or her needs. Moreover, CS relations often sustain unequal treatments of in-group and out-group members. Moral disagreement can occur if participants in an interaction employ different notions of fairness because they are framing the situation in terms of different relational models. For instance, compensations that would be appropriate within an MP model can cause offense when offered in domains constructed as CS relations. Monetary compensation in exchange for eschewal of certain sacred values (e.g., access to holy sites) can actually be counterproductive.[387]

Rai and Fiske attribute such a one-dimensional understanding of fairness as consisting merely in equality to both Haidt and Turiel. However, in Haidt's and his colleagues' recent work, the fairness foundation is about proportionality, while concerns for equality are rooted in concerns for care/harm and liberty/oppression.[388] According to Rai and Fiske, their theory relates to both Shweder's three ethics and Haidt et al.'s MFT by offering a social-relational framework that predicts when and how a specific ethic or foundation will govern moral judgment. Both the evaluation of harm and of what is fair depend on the relational model. Rai and Fiske state that the AR model enriches Haidt's authority foundation by emphasizing the responsibility of superiors towards subordinates. However, Haidt explicitly mentions such obligations.[389] Concerns with purity are, in Rai and Fiske's framework, typical of CS relations, where people strive for unity and avoid pollution of the relationship or the group's integrity. Unity violations often involve physical contact or incorporation. While Haidt et al. tie purity concerns to religion (not exclusively), Rai and Fiske

386 See ibid., p. 68.
387 See ibid., p. 66.
388 See Haidt (2012), pp. 182–183.
389 For instance in Haidt & Joseph (2007), p. 384.

believe that religion is important mainly in virtue of shaping the relations between individuals and supernatural entities or coreligionists.[390]

Emotions have a threefold function in relationship regulation: They gauge the "social-relational potential of others, generat[e] the desire to enter into social relationships with others, and regulat[e] existing social relationships […]."[391] Rai and Fiske speculate that disgust corresponds to violations of unity, while actions caused by compassion and empathy enhance unity. Pride is a feeling of entitlement linked to higher status or satisfactory role fulfillment in AR relations; respect and awe motivate obedience towards superiors. Gratitude is associated with reciprocal EM relations, although it seems to me that it might just as well occur in the context of MP and AR relations.[392] Rai and Fiske's claims regarding the relevance of their model are similar to Haidt's assessment of MFT: To deal productively with fundamental moral disagreement, it is necessary to realize that actions and evaluations with which one disagrees can, and often do, spring from moral motives. Stable consensus can result only from understanding the counterpart's mindset, particularly her use of relational models.[393] Rai and Fiske also touch upon what they think might be the moral-philosophical relevance of their approach; it is very much in the spirit of ought-implies-can arguments. We need to understand human nature and psychology if the prescriptions of normative ethics are to have the desired practical consequences. Moreover, relationship regulation enriches the vocabulary at our disposal to discuss ethical matters.[394]

390 See Rai & Fiske (2011), p. 67.
391 Ibid., p. 68.
392 See ibid.
393 See ibid., p. 69.
394 See ibid.

# 4 Emotions in Morality

Cognitive psychology, like behaviorism, used to be relatively indifferent to affective mental phenomena, or else saw them as producing irrational behavior.[395] From the 1980s onwards, however, there has been a surge of interest in emotions and their role in human cognition, an 'affective revolution'. Today, it is widely accepted that emotions are major determinants of human behavior, and that the output of affective processes need not be less valuable than that produced by more effortful, nonemotional cognitive mechanisms.[396] Evolutionary psychology adds plausibility to the apparent importance of emotional-intuitive processes: Adaptations need not produce optimal results in order to be superior to alternative designs. In addition, cognitive effort and slow processing are costly in terms of inclusive fitness.[397]

In order to understand the relation between emotions and morality, and moral relevance in particular, we need to know what an emotion is. Even though some debates regarding the conditions mental phenomena need to fulfill in order to count as emotions remain unresolved, chapter 4.1 provides a working definition for the remainder of this thesis.

## 4.1 What is an Emotion?

According to many definitions, emotions are constituted by characteristic patterns of deviations from a nonemotional baseline condition that are regularly elicited by the perception of particular situations or events (elicitors). Specifically, emotions are combinations of:

1) physiological change (arousal),
2) an appraisal (see below) or evaluation of stimuli (cognitive processes),
3) Qualia[398]: a feeling or phenomenological experience that is either positive/pleasant or negative/unpleasant to a certain degree (valence),
4) facial or motor expressions,
5) and specific action tendencies or motivations.[399]

---

395 See Scherer (2003), p. 565.
396 See Haidt (2003c), p. 852, Forgas (2003), pp. 596–597.
397 See Sunar (2009), pp. 450–451 and Verplaetse et al. (2009), p. 36.
398 See Rozin et al. (2008), p. 759.
399 See Zimbardo & Gerrig (2008), p. 731, Haidt (2003c), p. 853, Roeser (2011), p. 134. E.g., fear may involve restricted blood flow to the face and stomach (physiological change), conscious and unconscious interpretations of the dangerous situation (cognitive processes), a subjective feeling of being afraid (possibly the brain's response to the body's state of arousal), typical facial expressions, and a tendency to either fight or flee. Definitions vary in the emphasis they put on these phenomena.

Elicitors and physiological changes or action tendencies/motivations/behavioral expressions can be understood as input and output of (evolved) psychological mechanisms respectively (see section 1.2.3). However, emotions allow for more behavioral flexibility in comparison with other, more instinct-like processes: While lower organisms respond rather rigidly to stimuli, emotions provide behavioral *tendencies* that an organism does not necessarily act upon; both reconsideration of the eliciting event as well as of response alternatives are possible.[400]

Theories of emotion make different claims about the relation between the typical phenomenology of emotions and the corresponding physiological events.[401] According to the so-called James-Lange theory of emotion[402], the physiological response to an eliciting event is prior; the phenomenological experience is but the brain's response to the body's activity. The emotion *consists in* how the corresponding physiological changes feel. On this view, emotions do not necessarily contain *judgments* of any kind. It is therefore, in a specific psychological parlance, a 'noncognitive' theory of emotion.[403] The Cannon-Bard theory of emotion[404], in contrast, claims that the physiological response to a trigger causes activity in the autonomous nervous system and the phenomenological experience of an emotion in the brain *simultaneously*. This was suggested because some physiological responses (like a blush) seem to respond to triggers more slowly than the phenomenological experience (e.g., embarrassment) arises, because people frequently do not detect changes in their physiology, and because it seemed possible to elicit the same physiological changes usually caused by an emotion trigger *without* thereby inducing the emotional experience. Finally, Cannon believed that there are not enough different physiological activation patterns to match the diversity of emotional experiences. Both attention to this problem and the idea that emotional experience is the perception of one's physiological responses are incorporated in the two-factor theory of emotion proposed by Stanley Schachter and Jerome Singer: On that view, emotions are inferences about the *causes* of rather undifferentiated physiological responses (factor 1) from cues in the situation (factor 2).[405] Because bodily responses can be interpreted in various ways, the number of possible emotional experiences is larger than the number of physiological arousal patterns. Moreover, the emotion still is a perception of the

400 See Ellsworth & Scherer (2003), p. 572.
401 See Schacter et al. (2009), pp. 370–371.
402 Named after William James and Carl Lange.
403 See Prinz (2007b), p. 53. Note that in this context, 'noncognitive' just means 'not involving judgment'. There are still 'cognitive' processes going on in the sense that information is being processed.
404 Named after Walter Cannon and Philip Bard.
405 See Scherer (2003), p. 564, Zimbardo et al. (2006), p. 360.

physiological arousal, albeit one with more of a cognitive/interpretive component. Evidence indicates that people can indeed make mistakes in identifying the causes of their arousal, and that these mistakes can influence what emotions they experience, as well as their intensity. It turned out, however, that physiological responses vary more than Schachter and Singer allowed for in their notion of "undifferentiated physiological arousal"[406]. The current majority view appears to be that physiological arousal differs across emotions, and that emotional experience consists in the experience of that arousal to some degree (the heritage of the James-Lange theory), but is not mapped on it one-to-one. At least sometimes, assumptions as to the causes of the physiological activity do affect emotional experience.[407]

In a related debate on interpretive elements in emotions, researchers discuss whether emotions contain an evaluative judgment or "appraisal"[408] of the situation.[409] An appraisal is a sort of judgment about which elements of incoming sensory data are important; it is "an evaluation of the emotion-relevant aspects of a stimulus"[410], and it seems that the amygdala plays a major role in generating it.[411] Psychologist Richard Lazarus introduced the notion of 'core relational themes' that correspond to specific emotions and emphasized that emotions are essential to how we deal with others.[412] Such a theme is "a relation between organism and environment that occasions the onset of the emotion"[413], "the central relational harm or benefit in adaptational encounters that underlies each specific kind of emotion."[414] On Lazarus' view, emotions contain an appraisal of the respective relation. For instance, sadness contains an appraisal to the effect that there has been an irrevocable loss of some valued part of the organism's environment.[415] Appraisals can be generated along two routes, a 'fast pathway' along which impressions of a stimulus flow directly from the thalamus to the amygdala, and a 'slow pathway', along which such impressions pass through

406 See Schacter et al. (2009), p. 372.
407 See ibid., p. 373.
408 A term coined by Magda Arnold (1960): direct, immediate, intuitive evaluations. The term was used to account for qualitative distinctions among emotions. See Ellsworth & Scherer (2003), p. 572.
409 See Prinz & Nichols (2010), pp. 118–119.
410 Schacter et al. (2009), p. 374.
411 See ibid. Prinz (referring to Arnold, Lazarus, and Scherer) defines an appraisal as "a representation of an organism/environment relation that bears on well-being. Call such a relation a 'concern'." Prinz (2007b), p. 51.
412 See Ekman (2003), p. 24.
413 Prinz & Nichols (2010), p. 119.
414 Lazarus (1991), p. 121, quoted in Loewenstein & Lerner (2003), p. 628. Appraisals answer questions such as the following: Is this important to me? Do I understand what is going on? Can this be controlled? What caused it? Has a social norm been broken? See Ellsworth & Scherer (2003), p. 574. Different answers to these questions characterize different emotions.
415 See Ekman (2003), p. 83. Presumably, emotional intensity depends on the degree to which the thing in question was valued, and the perceived probability of its retrieval.

the cortex before reaching the amygdala.[416] The debate about the necessity of appraisals is sometimes presented as part of the question whether emotions are essentially cognitive, that is, whether they necessarily contain a 'cognitive' component.[417] The terms 'emotion' and 'cognition' frequently cause confusion:

> Cognition is often seen as an antagonist to emotion, as emotion is seen as an impediment to the proper functioning of the pinnacle of cognition—rational thought. This widely shared assumption is the result of a philosophical debate about the roles of passion and reason in human nature that goes back to Plato. In arguing for a tripartite structure of the soul, Plato created the concepts of "cognition," "emotion," and "conation" (motivation), and placed them in partial opposition to each other. […] The latest consequence of Plato's doctrine has been a debate on how much cognition is required for emotion, if any […].[418]

In line with an understanding of emotion and cognition as at least partially opposed, appraisal theorists have been accused of being overly cognitivistic: Appraisals seem to demand much more higher-order, elaborate and controlled information processing than what is actually necessary for emotions to occur. Appraisal theorists respond that these accusations rest on a misunderstanding: Although appraisals constitute evaluations of the significance of an event, they can be largely automatic and unconscious.[419]

On noncognitive views, emotions do not *necessarily* contain cognitive elements. Accordingly, experiences of sadness are possible without an appraisal to the effect that something valued has been lost.[420] Whether there really is a conflict between these positions depends, however, on what 'cognitive' means, or what appraisals actually consist in. If cognitive processes/appraisals can occur at very 'low', unconscious levels of processing, cognitive elements in or preceding many emotions become much less controversial.[421] Moreover, appraisals can be understood either as antecedents or as components of emotion. In the latter case, the boundary between cognition and emotion (or 'reason' and 'passion' in more traditional philosophical jargon), is blurred.[422]

---

416 See Schacter et al. (2009), p. 375.
417 See Prinz & Nichols (2010), pp. 118–119.
418 Scherer (2003), p. 563.
419 See ibid., p. 564.
420 See Prinz & Nichols (2010), p. 119.
421 "Leventhal and Scherer proposed the idea that appraisals can occur at three different levels, specifically the sensorimotor, the schematic, and the conceptual level, and that processes occurring at different levels can interact: Subcortical processes can stimulate cortical involvement and vice versa […]." Ellsworth & Scherer (2003), p. 576.
422 See ibid., p. 575.

There are some emotional phenomena, like emotional responses to music, which do not appear to depend on appraisals. Most of the emotional processes discussed below, however, do contain cognitive elements ('cognitive' in a modest sense); at the same time, emotions differ with respect to the various sorts of appraisals they require.[423] This is not an eccentric position. On the contrary, increasing effort is being devoted to integrating, instead of opposing, the concepts of emotion and cognition.[424] Even if appraisals are not necessarily conscious, emotions can still be characterized by core relational themes. We might find that sadness regularly occurs in response to environment-organism constellations that constitute losses, even though there is no conscious judgment that a loss has occurred.[425] Such observations, in combination with the finding that congenitally blind children show their emotions in facial expressions similar to those of sighted children, lend some support to the thesis that basic emotion themes are heritable, but subject to considerable modification through culture-specific learning.[426] In the following, I rely on a rather broad notion of emotion, according to which emotions *can* include appraisals of the situation in question (a weak noncognitive notion, if you will, as opposed to a strong noncognitive notion according to which emotions *never* contain appraisals). In fact, I believe many of the emotional responses involved in *moral* cognition do contain appraisals.

So far, we have seen that specific combinations of physiological changes, action tendencies, appraisals, and phenomenological experiences triggered by specific elicitors characterize different emotions. Some accounts, however, define emotions by reference to only some of these features. That can be expedient: For instance, a wide variety of eliciting events can entail anger. *Appraisal theories* emphasize the cognitive processes involved in emotions; their definitions of emotions revolve around specific combinations of appraisals.[427] *Dimensional* theories of emotion, in contrast, hold that emotions vary along two (sometimes more) dimensions and can be classified by the specific region they uniquely occupy in the space generated by these dimensions. Frequently, emotions are classified by pleasantness/valence (ranging from very pleasant to very unpleasant) and activation/arousal (ranging from high to zero arousal). Dimensional theories typically focus on the subjective experiential quality, or *qualia* of emotions. In contrast to *categorical* theories of emotion, which posit a limited number of clearly distinct emotions or emotion families, dimensional accounts, like ap-

423 See Scherer (2003), p. 564.
424 See ibid., p. 563.
425 See Prinz & Nichols (2010), p. 119.
426 See Ekman (2003), p. 26.
427 See Ellsworth & Scherer (2003), p. 586.

praisal theories, allow for infinite numbers of emotions, characterized by all possible combinations of values on the dimensions under consideration.[428] Dimensional theories often merely describe the emotions, but do not explain their adaptive function. Appraisal theorists, however, consider at least the two-dimensional approach inadequate to attain this descriptive goal: "[F]ear and anger cannot be distinguished simply on the basis of differences in levels of activation and pleasantness. […] [W]e need to know more about how the organism interprets its situation."[429] In sum, emotion theories emphasize various important characteristics; many of them help to understand the various relations between emotions and morality.

## 4.2 Defending Cognitive Theories of Emotions

Philosopher Jesse Prinz argues that strong emotionism, a position that combines epistemic emotionism (the view that "moral *concepts* are essentially related to emotions"[430] and metaphysical emotionism ("moral *properties* are essentially related to emotions"[431]), is hard to reconcile with cognitive theories of emotions. Cognitive theories of emotions hold either that emotions *consist in* cognitive states like judgments or appraisals, or that they necessarily *contain* such cognitive elements. According to him, combining cognitive theories of emotion with emotionism implies that moral judgments contain moral emotions (emotionism), and that these moral emotions in turn contain moral judgments (the cognitive aspect).[432] This is, I think, neither problematic nor true. Prinz mentions guilt as an example: If guilt is cognitive, does it contain the appraisal that I have done something *morally* wrong? Not necessarily so. It seems possible that people experience guilt merely because they believe they have harmed somebody more or less close to them. Such an experience is conceivable without a notion of *moral* wrongness that would require more than the aforementioned belief, a corresponding emotional response of negative valence, and a motivation to make amends (see chapter 4.5.1.2). Prinz attempts to refute this response with an open-question argument in the spirit of G. E. Moore: If wrongness contains the experience of guilt (because moral judgments consist in emotions), and guilt contains the appraisal that somebody has been harmed (because emotions contain judgments/appraisals), then 'harming is wrong' just

428 See Haidt (2003c), p. 855, Ellsworth & Scherer (2003), p. 574, Schacter et al. (2009), p. 369.
429 Ellsworth & Scherer (2003), p. 574.
430 Prinz (2007b), p. 14, my emphasis.
431 Ibid., p. 16, my emphasis.
432 See ibid., pp. 54–55.

means 'harming is harming', which is uninformative. However, identifications of moral concepts, like wrongness, with nonmoral facts frequently *surprise* us. We have the impression that it does *not* betray confusion about moral wrongness to ask, "This act/situation is an instance of nonmoral fact X, but is it *really* wrong?"; the question remains open. According to Prinz, such an argument can be leveled against any emotionist position committed to a cognitive theory of emotions.[433]

Since I employ a concept of emotion according to which emotions *can* contain appraisals, and will argue that emotions establish fundamental moral concerns, I should comment on this argument. The first thing to note is that, if the mental capacities involved in moral judgment are to some extent modularized, the claim that moral wrongness is a unified concept which *always* contains guilt is dubious. Attributions of moral wrongness might well rest on several distinct emotional experiences of negative valence. Does that suffice to refute Prinz's argument? Maybe not. If there were different variants of moral wrongness, each of which contains a specific emotional experience (anger, guilt, disgust, etc.), one might still ask why people can be surprised by the claim that this particular kind of wrongness contains that specific emotion, which in turn contains specific nonmoral appraisals. If that specific kind of wrongness *really* contained that specific kind of appraisal, impressions of having received unexpected information should not occur. Prinz's claim should be criticized differently: Firstly, emotions are not made up only of appraisals, but also of a certain 'feel', an action tendency, etc. If we include *all* these components, a given moral concept would not be equated with just the appraisal, but with the *combination* of all these aspects of emotions. Such a more complete account of moral concepts might seem less surprising, and accordingly, the corresponding question less open.

On a more fundamental level, I believe that there is good reason to question the relevance of the fact that people are 'surprised' at the explanation of moral concepts in terms of psychological concepts with regard to the adequacy of such explanations. These concepts are not the kind of input the processing of which will generate the kinds of experiences which usually mark moral evaluations. Therefore, an emotionist perspective can be combined with a cognitive understanding of emotions.

Prinz argues not only that cognitive theories of emotion are in tension with emotionism, but also that noncognitive theories are *independently* more plausible than cognitive theories. Prinz defines a minimum requirement of what an emotion would have to contain in order to count as cognitive and argues that, given this minimum, cognitive states are not necessary

[433] See ibid., p. 55.

for emotions. At a minimum, he claims, cognitive states contain concepts, and concepts are mental representations that can be (freely) combined with other such representations. Given this understanding of the cognition requirement, Prinz claims that cognitive theories are implausible for several reasons. Firstly, the minimum requires more than is and was present when emotional phenomena emerge(d) in both ontogeny and phylogeny. The minimum also appears to be incompatible with the immediacy of many emotional experiences.[434] Moreover, it is possible to elicit some emotions without appraisals, for instance if generating the physical correlates of an emotional experience, such as a smile, elicit the corresponding emotional experience (facial-feedback hypothesis).[435] Drug use is another example, although one might question to which extent the emotional effects of various drugs are independent of changes in appraisals. In addition, research on the two pathways of emotion processing suggests that at least some emotions can arise without involvement of the neocortex, where one might think appraisals/judgments occur. Of these, at least the facial feedback example is a good argument to show that emotional experiences do not require appraisals. From an evolutionary perspective, however, the function of emotions can only be understood if the physical responses and action tendencies are considered in combination with the adaptive elicitors and appraisals. Moreover, all of these arguments hinge on what exactly appraisals are. If one believes that Prinz's minimal conception of cognitive states is too demanding his arguments lose their grip. Consider for instance the fear response to the visual perception of a coiled snake or threatening facial expressions that can arise in subjects with lesions in the visual cortex via the thalamo-amygdala pathway.[436] If an appraisal theorist argues that there has to be an appraisal, conscious or not, of a lengthy moving object or fearsome face at some point in the processing chain in order for that emotion to occur, the conflict seems to dissolve: Prinz argues that it "would be totally untenable to claim that the thalamus or the amygdala harbor concepts."[437] However, this untenability depends on his definition of concepts as freely combinable representations. If representations do not have to be conscious or 'freely combinable' (whatever that means), his criticism is less convincing. The appraisal theorist can argue that even in the case of thalamo-amygdala processing, there has to be an appraisal of the snake-like shape or the facial expression at some point in the processing pathway for these responses to occur. The

434 See ibid., pp. 56–57.
435 See De Waal (2013), p. 132.
436 See Prinz (2007b), p. 57.
437 Ibid.

noncognitivist might respond that such a notion of appraisals is uninformative. Does the appraisal theorist have a rejoinder?

Prinz defends a theory of embodied emotions in the spirit of the James-Lange account, according to which somatic states are necessary and sufficient for emotions to occur. In response to an objection such a position could face, namely that it cannot account for the possibility of inappropriate emotions, Prinz refines his concept of emotion. In my view, this refinement makes it quite similar to cognitive theories of emotion that employ a less demanding notion of cognitive states or appraisals than the one Prinz defined in order to reject cognitive theories of emotion: He claims that impressions of inappropriateness require that emotions *represent* something, and that they can *mis*represent that something. While cognitive states or judgments of the kind that Prinz considers to be overly demanding can accomplish representation, he now suggests that emotions can be noncognitive states with representational content. What does representation mean? Prinz considers this the most promising account of representation:

> [A] mental representation, M, represents that which it has the function of reliably detecting. Roughly, M represents that which it was set up to be set off by. […] A beep [of a smoke alarm] represents smoke, because it is reliably caused by smoke and it [the smoke alarm] was engineered so as to be caused by smoke.[438]

In order to apply this notion of representation to emotions, we have to find out what reliably elicits certain emotions, and by what they were 'designed' to be elicited. In endorsing the account of representation quoted, Prinz deviates from the James-Lange account. James and Lange thought that bodily changes not only cause emotions, but that emotions also *represent* these changes. On Prinz's view however, emotions represent organism-environment relations that are important; they represent "concerns"[439]. Prinz argues that the various representations that can trigger (specific) emotions are not themselves parts of the emotion, but rather causes of these emotions.[440] He refers to his amendment of the James-Lange theory of emotion as 'embodied appraisal theory'.

There is a second criticism of noncognitive emotion theories that we already touched upon earlier. Are there enough distinct bodily states to account for all the different emotions we experience? My own impression is that the criticism is void if brain states count as bodily states. Prinz, however, considers this a serious problem. In response, he claims that we can

438 Ibid., p. 61.
439 Ibid., p. 63.
440 See ibid., pp. 63–64.

distinguish emotions whose corresponding bodily states are very similar by the kind of appraisal that caused them, and these kinds, in turn, by what they represent.[441] I believe that this move does in fact commit him to understanding appraisals as *parts* of emotions. In sum, I believe that it is possible to maintain that emotions can contain appraisals and defend a view according to which emotions are essential to moral judgment.

## 4.3 Moral Emotions: Present Theories and a New Proposal

What motivated the preceding close look at emotions? Several studies in moral psychology uncovered aspects of moral judgment that might undermine its trustworthiness. This suspicion is frequently based on the perceived moral relevance or irrelevance of the 'factors' thought to affect moral judgment according to newly discovered properties of the judgment process (evolutionary explicability, intuitiveness, emotional influence, heuristic mechanisms, etc.). Some authors claim that these findings can help adjudicate among conflicting judgments and thereby advance the moral debate. However, they rarely scrutinize their impressions of moral relevance. To avoid stalemates between intuitions about moral relevance that are just as unresolvable as those between first-order intuitions are, and to evaluate the merit of relevance-based debunking arguments in general, I surveyed psychological accounts of moral judgment and of the perception of moral relevance. The moral/conventional framework descended from the cognitive-developmental tradition; it emphasized cognitive, as opposed to emotional processing and concerns with harm, rights and justice. Cultural and evolutionary psychology, instantiated in moral foundations theory, attribute more significance to emotions in both general and moral decision making. Assuming that moral foundations theory more adequately captures morality around the globe than the moral/conventional tradition does, we still need to know more about how these 'fundamental concerns' come to appear morally relevant. Both an appropriate assessment of whether emotional influence on moral judgment is a bias or not, and an understanding of relevance assessments and their reliability require an account of the relations between emotions and moral judgment.

According to MFT, emotions function like moral taste buds that detect specific flavors of relevance. However, the exact connection between emotions and morality has thus far only been hinted at. Both psychological and philosophical literature speak of 'moral emotions'. In this chapter, I discuss characteristics in virtue of which emotions are considered

[441] See ibid., pp. 66–67.

'moral', and develop a categorization of these characteristics that I believe captures the connections between emotions and moral relevance more comprehensively. I also describe morality-related functions of several emotions.

The most important distinction I want to make among the ways in which emotions relate to morality is between the *foundational* relation and the *instrumental* relation. Moral foundations theory exemplifies the *foundational* relation:[442] Emotions link specific social events or situations (input) to specific mental and behavioral output. Certain classes of input motivate people to modify their behavior in broadly similar ways across cultures. These connections between certain types of elicitors or appraisals and action tendencies (or the preparedness to learn these connections more easily than others) were, presumably, to some extent shaped by evolutionary processes. The respective emotions can be triggered not only if relevant events happen to the self, but also if others are affected. In my view, the foundational relation between emotions and morality is essential for an understanding of moral relevance: Triggers that we respond to emotionally for evolutionary reasons are, I claim, more likely to count as morally relevant across generations and cultures than those without the backing of emotional EPMs. Moral relevance is an appraisal of importance contained in emotional responses to (at least partially) evolutionarily selected types of elicitors, where elicitors and/or action tendencies can be (but are not necessarily) tied to the well-being of others. Without an emotion-based ability to not only perceive harm and well-being, hierarchical relations, reciprocity, distribution of resources, purity, etc., but also to care about and act because of these perceptions, we would not consider these things morally relevant. Critics might remark that humans also respond to these triggers after rational deliberation. However, why would they deliberate at all if they did not care in the first place? Moreover, a nonemotional notion of moral relevance is hard to reconcile with the increasing amount of empirical evidence pointing to the pervasive effect of emotions in moral judgment. It seems that we construct the domain of morality from emotional sensitivities for fundamental concerns.

Foundational influence is not the only way in which emotions relate to morality. Several authors distinguish 'moral' from nonmoral emotions, and their criteria often concern whether and how emotions promote *morally desirable outcomes* (instrumental relations). For instance, transgressions of moral norms or praiseworthy behaviors trigger specific emotions. Emotions occur regularly when we pass moral judgment, and they motivate moral

[442] Note that individual emotions can relate to morality both foundationally and instrumentally. I will illustrate the various effects in more detail in chapters 4.4 and 4.5.

action.[443] Note that the classification of a given emotion as instrumentally moral presupposes the existence of a notion of the morally desirable and undesirable (i.e., the morally relevant), which is itself rooted in emotional processes.

### *4.3.1 Haidt on Moral Emotions*

Jonathan Haidt proposes a two-dimensional characterization of emotions as moral that focuses on elicitors and action tendencies. In contrast to other dimensional theories, his account does *not* sort emotions according to experiential criteria like 'intensity of arousal' and pleasantness/unpleasantness, but rather according to features of actions (both elicitor and response). An emotion is prototypically moral if it is 1) easily triggered by events that affect the interests of others or society (which does not preclude that they can also easily be triggered on one's own behalf) and 2) contains 'prosocial' action tendencies, where prosocial means either 'benefitting others' or 'stabilizing a given social order', including punishment of deviant behavior.[444]

Haidt's moral emotions are not primarily self-interested; they can even motivate action that is costly to the agent. These emotions enable social existence by preventing individuals from advancing their own interest in a destructive fashion, both through their motivational effects and their ability to convey intentions.[445] While disinterested elicitors and prosocial action tendencies mark prototypical moral emotions, emotions in which either of these characteristics is less pronounced are considered less 'moral' as a matter of degree.[446] Many moral emotions are responses to deviations from implicit and explicit social norms and stereotypes.[447]

---

[443] See Haidt (2003c), p. 853.

[444] See ibid., pp. 854–855. There, Haidt also writes: "An alternative definition of the moral emotions can […] be stated as the difference between the emotional life of Homo sapiens and the emotional life of […] a perfectly selfish creature, […] who cares only about her own well-being and who cooperates with others only to the extent that she expects a positive net payoff from the transaction." He equates the "perfectly selfish creature" with *homo economicus*. However, the homo-economicus model allows that the well-being of others be part of an agent's utility function, which makes it difficult to say whether behavior that benefits others is selfish or not. Haidt seems to think that such integration of the interests of others is incompatible with the homo-economicus model, thus I take him to mean that the perfectly selfish creature's utility is unaffected by the well-being of others. Note that Haidt's definition refers to the *emotional life* rather than the actions of an agent. Nevertheless, the difficulty is not eliminated, but rather moved to the level of mental states: If I am motivated to punish free riders who I'm not going to interact with again, is my motivation really other-interested, or is it just something that human beings can't help but feel, so that they experience negative affect if punishment is suspended? For a helpful discussion of psychological egoism vs. altruism, see Stich et al. (2010).

[445] See Zimbardo et al. (2006), p. 351.

[446] See Haidt (2003c), pp. 853–854.

[447] See Moll et al. (2008b), p. 6.

### 4.3.2 Prinz and Nichols on Moral Emotions

Jesse Prinz and Shaun Nichols take a slightly different approach to identifying moral emotions. The first criterion they consider is association with moral rules as specified in the moral/conventional framework. This is unsatisfactory for reasons already mentioned: The moral/conventional tradition conceives of moral rules as independent of the opinions and practices of others (authority independence). However, norms such as "Follow the ways of your elders!" or "Obey the dietary customs of your society!" do not seem to fulfill that condition, but count as moral in some cultures. (Unfortunately, Prinz and Nichols do not state how they determine that rules count as moral in non-Western cultures.) On the moral/conventional view, the status of a rule depends partly on whether it is justified with or without reference to opinions and practices of authorities, but children seem to distinguish between moral and conventional rules even though they are not much concerned with *how* the rule is justified. Moreover, authority independence is not exclusive to moral norms: Some nonmoral norms also do *not* depend on social convention (personal, prudential norms). On the other hand, if relativism is true, *all* rules depend on convention, but relativists might still want to call only *some* rules 'moral'.[448]

In a second attempt to describe the relation between emotions and morality, Prinz and Nichols stipulate which emotions are moral and delineate the moral domain correspondingly.[449] Only norms adherence to or violation of which elicit this subset of emotions count as moral. Indeed, in Western transgressors, the violation of prototypically *moral* rules tends to trigger shame or guilt, while violations of prototypically *conventional* rules are associated with embarrassment. Unfortunately, this procedure seems circular: The aim was to "define moral emotions in terms of moral norms, and now moral norms [are being defined] in terms of moral emotions."[450] In light of these difficulties, Prinz and Nichols define moral emotions as associated with paradigmatic moral rules. 'Association' takes the shape of two major roles for emotions: as sources of motivation to act morally, and as determinants of moral judgments. The motivational role suggests itself since emotions are important sources of motivation, not just in moral matters.[451] Prinz and Nichols propose two types of motivational effects: In "judgment motivation", moral judgment precedes *actions*. In these cases, agents have a conscious notion of what they think morality demands, and emotions generate their motivation to act accordingly. However, moral judgments do not seem *necessary* for

448 See Prinz & Nichols (2010), pp. 120–121.
449 See ibid., p. 121.
450 See ibid.
451 See ibid., p. 112.

emotional motivation to behave in accordance with morality. A lot of research investigates how emotions affect the motivation to behave prosocially or cooperatively. In these cases, and to the extent that prosocial, cooperative, or other behavior counts as moral, emotions might generate motivation to act in accordance with the demands of morality even though the agent does not consciously consider what morality demands.[452]

Regarding the relation between emotions and moral *judgment*, several results suggest that emotions can function as moral intuitions, which determine or even constitute moral judgments. Haidt believes that such intuitions dominate most judgments, while others, like Greene, think that this is true only for some of them (see chapter 5). (However, Greene et al.'s data do not show that emotions are *not* active in consequentialist judgments.) Most fMRI studies find emotional activation when subjects engage in moral judgment.[453]

### 4.3.3 Valdesolo and DeSteno on Moral Emotions

Valdesolo and DeSteno, inspired by Adam Smith, have identified yet another connection between emotions and morality: Even emotions that are of a more self-interested nature, i.e., which are neither in their elicitors nor in their *proximate* action tendencies related to the 'interests of society or other people', such as jealousy, vengeance, or pride, may ultimately contribute to collective well-being. Thus, to the extent that the promotion of general well-being is a moral goal, even some emotions commonly considered vices might count as 'moral emotions'. In this case, emotions are 'moral' in virtue of the *outcomes* the respective action tendencies produce, not in virtue of the character of the agent's *intentions*. The moral status of outcomes depends on preexistent moral standard (which, as I have suggested, might be shaped by emotions in the first place). While my description of the effects in question is slightly more explicit than what Valdesolo and DeSteno offer, it appears to fit their examples: They classify jealousy and revenge as moral emotions since they "contribute to the evolution of flourishing cooperative societies."[454] On their account, 'pride' might also be 'moral', since it motivates people to compete and be better than others, thus promoting progress that supposedly benefits society as a whole.

### 4.3.4 Horberg, Oveis, and Keltner on Moral Emotions

Horberg et al. classify moral emotions according to the appraisals contained in them and the moral judgments to which they relate. They argue that specific appraisals contained in

452 Ibid., p. 113.
453 See ibid., pp. 114–115.
454 Valdesolo & DeSteno (2011), p. 276.

various emotions that affect moral judgment direct our attention to different "sociomoral concerns"[455]. On their view, emotions "influence moral judgment […] through core appraisals that are semantically related to a specific sociomoral concern (e.g., purity) and that remain salient throughout the emotion"[456], rather than affect judgments about all kinds of concerns. They cite research that indicates domain-specific effects of emotions on moral judgments: Disgust does not render moral judgments more severe in all areas of concern, but mainly affects judgments about issues related to purity concerns, even when disgust was elicited by nonmoral events.[457] In individuals with a high need for cognition, anger has similar effects with respect to justice concerns. In addition to such *domain-specific* effects of emotion on moral judgment and behavior, Horberg et al. posit *emotion-specific* effects on moral judgment. For instance, they expect that anger would influence moral judgments related to justice concerns, but that disgust or fear (which also have negative valence) would not. What is the difference between these effects? *Domain specificity* predicts that specific emotions affect only moral judgments that deal with specific sociomoral concerns, rather than all kinds of moral judgments. *Emotion specificity* predicts that only specific, but not all emotions affect moral judgment with respect to a given sociomoral concern. In Horberg et al.'s view, emotion specificity indicates that not only *valence,* but also the different *appraisals* contained in emotions affect moral judgment.[458]

They also discuss so-called embodiment effects. Subjects induced to mimic the bodily marks of specific emotions display effects on judgment and cognition that go along with these emotions when they occur 'naturally'. Most important for my project, however, is their discussion of how emotions relate to the process by which different issues "attain moral significance within a particular society or generation."[459] *Moralization* is the process by which specific moral judgments are integrated into a system of values. Horberg et al. mention changes that occurred in attitudes towards smoking and meat eating as examples. Are all issues equally likely to be moralized? Supposedly, in the U.S., issues are likely to be moralized when framed in terms of suffering or unfair treatment, a hypothesis that fits well with Haidt's characterization of Western cultures as focused on the care/harm and fairness/cheating foundations respectively.[460] Similarly, differences in the prevalence of disgust

455 Horberg et al. (2011), p. 238.
456 Ibid., p. 239.
457 See ibid.
458 See ibid., p. 240.
459 Ibid., p. 241.
460 See ibid., p. 242.

when confronted with issues such as abortion or gay marriage might explain the differences in the moralization of these issues between liberals and conservatives.

### 4.3.5 Foundational and Instrumental Moral Emotions

How can these connections between emotions and morality be categorized? On the one hand side, there are *foundational* effects, by which emotional mechanisms and their propensity to respond to specific kinds of elicitors, in combination with cultural processes that determine which phenomena are seen as belonging to these specific kinds, shape the subject matter of morality everywhere. The *instrumental* relations between emotions and morality rely on the notions of morality that were thus shaped by emotional mental processes. Among instrumental relations, the following distinctions might be helpful: On the instrumental perspective, emotions are moral because their action tendencies tend to maintain or bring about states of affairs that are morally commendable according to some given notion of morality. We should distinguish between cases in which emotions make us *intend* these commendable outcomes, and cases in which emotions generate actions that cause commendable outcomes as *un*intended side effects. Valdesolo and DeSteno's argument for considering jealousy, vengeance, or pride as moral emotions might fit the latter category. Emotions that make us *intend* morally commendable outcomes can be distinguished more finely. I propose a fourfold categorization by combining two binary criteria, namely 1) whether emotion and action involve an understanding of 'what morality requires', and 2) whether the motivation to act is intrinsic or extrinsic. Intrinsic motivation generates actions that are themselves rewarding, while extrinsic motivation causes actions that are not themselves rewarding but that promise an 'external' reward.[461] Thus, emotions can motivate to bring about morally commendable outcomes based on a moral judgment intrinsically if, for instance, acting in accordance with a moral code makes us proud or happy just for the sake of conformity to the code. On the other hand, intending to bring about morally commendable outcomes can also be based on extrinsic motivation, for instance if we act in accordance with a moral code because we fear punishment handed out to nonconformists, or enjoy the praise that comes with doing what is considered good in one's culture.[462] Admittedly, it might not always be possible to keep these mechanisms apart. Happiness elicited by doing what morality requires, for instance, might not arise in people (even if nobody

[461] See Schacter et al. (2009), p. 397.

[462] Batson (2011) introduces a similar distinction between *moral hypocrisy* (related to extrinsic motivation) and *moral integrity* (related to intrinsic motivation).

knows about their praiseworthy behavior) had they not at some point experienced and internalized the joys of being praised for doing what one ought morally to do. However, the idea is that the external motivation can lead even those who do not enjoy acting morally or suffer from pangs of conscience to obey moral norms nevertheless, in order to obtain some (other) kind of reward or avoid punishment.

What about cases in which intrinsic or extrinsic motivation to do what is morally commendable do not depend on awareness of a moral code, but in which the agent nevertheless (unknowingly) intends the outcome that is morally praiseworthy? Let me clarify that categorizing actions in this way does depend on an established notion of what morality requires. What distinguishes this category from the two constellations mentioned above is that no understanding of what morality requires need exist *in the agent* for this kind of emotion-induced motivation to occur. Let us first consider intrinsic, nonmoral motivation to do what morality requires. I have in mind, for instance, helping behavior motivated by immediate sympathy or empathy. In the case of extrinsic motivation, in contrast, such behavior is instrumental to achieving some other goal (access to social contacts, partners, resources, support; avoidance of aggression, etc.). Note that these nonmoral motives can plausibly be attributed to nonhuman primates and some other social animals, while both intrinsic and extrinsic *moral* motivation mediated by emotions seems uniquely human. The intrinsic premoral emotional motives I have in mind are closely related to those emotional mechanisms that have the foundational effect on morality touched upon above: Mechanisms that contain concerns for harm, care, hierarchy, purity, loyalty, etc. shape what is considered relevant in culturally amplified notions of morality. Thus, it is likely that action tendencies stemming from these mechanisms are commendable according to moral codes they helped establish. It is also conceivable that culture affects intrinsic motivation to produce morally commendable outcomes that does *not* involve awareness of a moral code. Classification sometimes is difficult. A child whose disposition to treat unknown in-group members kindly has been molded to extend only to people of the same complexion might be said to be intrinsically, but nonmorally motivated to conform to her learned behavioral tendencies. On the other hand, these cultural influences constitute rudimentary moral codes, so that is hard to say whether this particular motivation is moral or nonmoral.

This categorization of relations between emotions and morality is not exhaustive. Emotional motivation can be conscious or unconscious;[463] it can be "approach motivation,

[463] See Schacter et al. (2009), p. 399.

which is *a motivation to experience a positive outcome*"[464] or "avoidance motivation, which is *a motivation not to experience a negative outcome*."[465] Moreover, emotions directed at oneself differ from those directed at others, or depending on whether the *patient* of a triggering behavior is the self or a third person. The *object of the emotion* (self/other) is the guiding distinction in the discussion of individual emotions in chapters 4.4 and 4.5, while positive and negative *valence* provide a second level of categorization. For instance, *other-conscious* emotions can be negative (other-critical) or positive (other-praising). Beyond this division, reactions to others often involve further emotional phenomena that manifest sensitivity to their well-being (other-suffering). This last category (other-suffering) comprises sympathy, empathy, and compassion. While it is debatable whether these are indeed emotions (see chapter 4.4.3), I address them as such since they are closely tied to emotional phenomena and central to morality. I refer to emotions whose object is the self as *self-conscious*. Like other-conscious emotions, they can be classed according to positive (self-praising) or negative (self-critical) valence.[466] Haidt believes that other-critical, other-suffering and self-critical emotions are particularly important for moral behavior.[467] The following sections provide an overview of emotions central to each of these classes, focusing on elicitors, action tendencies, and the evolutionary challenges that shaped them. Since I am interested in moral relevance, I focus on emotions at the base of concerns that frequently count as moral. Keeping the characterization of fundamental concerns as learning modules in mind, I concentrate on mechanisms related to concerns recognizable across many cultures (i.e., the presence of a concern for *some* conception of fairness, rather than specific fairness norms).

## 4.4 Other-Conscious Moral Emotions

This section discusses emotions which arise in response to acts done by others, and are directed at those others. Among these, I distinguish between emotions of positive and negative valence (other-praising and other-critical, respectively). I will also address important emotion-related capacities such as sympathy or empathy, even though they do not, strictly speaking, fit the definition of emotions developed in chapter 4.1. They are nevertheless frequently referred to as 'moral emotions', and certainly significant for moral cognition.

464 Ibid., p. 400.
465 Ibid.
466 The terminology is adopted from Moll et al. (2008b).
467 See Haidt (2003c).

### 4.4.1 Other-Critical Emotions

Anger, contempt, and disgust are prominent *other-critical* emotions. According to Paul Ekman's categorization, they are *basic* emotions: each is associated with distinct, cross-culturally recognized facial expressions and specific physiological changes.[468] Nevertheless, there has been some debate about how exactly they differ, and some authors suspect that they are not fully distinct.[469] Among accounts that consider contempt, anger, and disgust separate psychological phenomena (fully differentiated models), the so-called CAD-triad hypothesis is particularly popular: It suggests that these three emotions each arise in response to violations of a specific moral code, as defined by Richard Shweder. Community violations elicit contempt (C), transgressions of autonomy norms arouse anger (A), and disgust corresponds to the domain of divinity (D) (violations of the natural order).[470] Another fully differentiated model proposed by Hutcherson and Gross states that the eliciting events of these emotions differ in the extent to which they require immediate action, as well as concern with intentions or competence.[471] I will address these accounts as I discuss each emotion in detail. Both the CAD-triad hypothesis and the model proposed by Hutcherson and Gross share a functionalist perspective: Emotions are evolved psychological mechanisms constituted by combinations of appraisals, physiological responses, communicative gestures, and action tendencies that developed to solve adaptive problems in the environment of our evolutionary ancestors.[472]

At least with respect to eliciting events, action tendencies, and some physiological changes, phenomena very similar to human anger can be observed in nonhuman primates when they interact with conspecifics. As for contempt, there is some debate regarding whether it is a separate emotion, with distinct protoforms in nonhuman primates. Disgust is probably not elicited by appraisals of social situations in nonhuman animals. What appears to distinguish the emotional capacities of humans more clearly from those of their nonhuman relatives is that we also experience these emotions when we watch *others* interact, even though the elicitation of anger seems to be stronger when the act in question has higher self-relevance. Other-critical emotions are typically responses to violations of explicit

468 See Sunar (2009), p. 153. There are many proposals for primary or basic emotions. See TenHouten (2009), p. 14 for an overview.
469 See Hutcherson & Gross (2011), p. 719.
470 See Rozin et al. (1999), Hutcherson & Gross (2011), p. 720, Prinz & Nichols (2010), p. 122, Looren de Jong (2011), p. 122.
471 See Hutcherson & Gross (2011), p. 734.
472 See ibid., p. 720.

or implicit norms and affect one's relationship with the perpetrator. These functional characteristics point to evolutionary rationales for each other-critical emotion: By virtue of natural selection, animals are endowed with motives or action tendencies that serve their own interest. Self-interested motives are likely to generate conflict in social species. Thus, reaping the inclusive-fitness benefits of social living requires bounds on self-interested behavior to ensure that the cost of conflict does not overcompensate the benefits of cooperation. Moral emotions are often elicited by acts that either endanger or promote the functionality of this organization, and create motives to behave in group-stabilizing ways.[473]

#### *4.4.1.1 Anger*

Anger is often considered one of the 'basic' emotions marked by relatively sharply defined elicitors, physiological responses (including facial expressions), and action tendencies.[474] With respect to these criteria, it seems that many other mammals besides humans are capable of experiences very similar to anger.[475] This observation indicates that the capacity for anger is an evolved psychological mechanism. Outrage triggered by violations of moral norms probably evolved from processes elicited by goal blockage and frustration in more primitive organisms. While such events still cause anger in humans, *moral* anger[476] is typically associated with appraisals of harm or insults, injustice or unfairness, and self-relevance. Apart from self-relevance, these appraisals fit quite well with violations of norms belonging to the ethics of autonomy (care/harm & fairness/cheating).[477] It can be felt on behalf of ourselves or others we identify with (personal anger) and on behalf of others we care about (empathic anger). Moreover, it can be triggered by the appraisal that some standard has been violated (principled anger).[478] These variants of anger differ slightly both in the appraisals which elicit them and their motivational consequences: Principled anger results

473 See Hynes (2008), p. 27.

474 See Prinz & Nichols (2010), p. 124. Ekman remarks that there might actually be several themes for anger. See Ekman (2003), p. 110.

475 The characteristic experience (qualia) of these phenomena in animals is hard to assess, therefore it is not part of the comparison.

476 I use 'anger' as umbrella-term that comprises 'outrage' and 'indignation'.

477 See Tangney et al. (2007), p. 361, Prinz & Nichols (2010), p. 125. Anger is also elicited by many other occurrences which are not easily captured in terms of injustice (e.g., annoying behaviors of others), but can nevertheless be subsumed under the theme of autonomy violations. "[I]f anger (or a close analogue) reaches deep into our mammalian ancestry, it would be surprising if it turned out that the only activator for anger is an appraisal of unfairness. In older phyla, the homologues of anger may be more typically elicited by physical attacks by conspecifics or the taking of resources (battery or theft). Similar responses may also arise in hierarchical species, when an animal that is not dominant tries to engage in dominant behavior or take a privilege that is reserved for dominant individuals. […] Being annoying or disruptive, thwarting goals, violating personal possessions of space, being insulting or offensive—all these things have a negative impact on a victim, and thus fail to respect individual rights or autonomy. Injustice is just one special and prevalent case." Ibid., pp. 129–130. See also Hutcherson & Gross (2011), p. 733.

478 See Batson (2011), p. 233.

from an appraisal of norm transgressing, independent of whether any interests of the self, friends, or strangers are affected, and motivates reestablishment of the norm. It can be understood as an example of intrinsic motivation to bring about a morally commendable outcome that rests on an understanding of what morality requires. Personal anger, in contrast, is elicited if the self or its in-group have been (undeservedly) harmed and motivates to "undo the harm and/or punish the harm-doer"[479]; such anger might occur with or without being based on awareness of a moral code. In the case of empathic anger, the motivation is similar in its effects (repair and/or punish), but it is altruistic since it is elicited by acts done to others, and tends to serve the victims' wellbeing rather than the self.[480]

I suggest that anger is the foundation of some moral concerns in the sense that regular angry responses to, for instance, *physical harm* done to the self and in-group members lead to the establishment of informal and, in time, formal norms prohibiting harmful behavior. Physical harm is morally relevant not only, but also because the appraisal of such harm is a prototypical anger elicitor. One might object that there is a moral concern more basic than anger, namely one's own or another's well-being. We do care about harm done to others and ourselves even if there is nobody to blame for it, as in case of illness, accidents, or natural disasters. Section 4.4.3 on other-suffering 'emotions' addresses these issues. However, with respect to moral judgments, anger is significant in its own right. We care about *agents* and the evaluation of their behavior. Events caused by 'agents' who do not understand norms cannot be governed by norms.[481]

The capacity to respond with anger can also be recruited by norms whose object of protection is not, as in the case of freedom from physical harm, deeply embedded in evolved emotional responses, but rather the product of cultural processes. Anger generally motivates aggressive behavior towards norm violators/aggressors and can contribute to the reestablishment of the social order in various ways.[482]

> [T]he angry person reports becoming stronger…in order to fight or rail against the cause of anger. His or her responses seem designed to rectify injustice—to reassert

[479] Ibid.

[480] See ibid. From an evolutionary perspective, it makes sense to think of different anger-induced action tendencies as coupled to specific themes or elicitors. See Ekman (2003), p. 112.

[481] See Fehr & Gächter (2002) on the importance of negative emotions for altruistic punishment and cooperation.

[482] See Prinz & Nichols (2010), p. 124. The discussion is cursory. Here, as with regard to all the other examples given in the following, it easy to imagine much more fine-grained and more comprehensive accounts of emotions, elicitors, and action tendencies. Responses to injustices, harm done, or unfairness, for example, could consist in sophisticated plots of vengeance or immediate, physical attack.

> power or status, to frighten the offending person into compliance, to restore a desired state of affairs.[483]

These behavioral effects do *not* require that the punitive impulses stem from an explicit intention to restore order or improve society. To some extent, people simply respond to perceived harm with a motivation to retaliate (often in a way that is similar to the original offense); and they do so even if a peaceful resolution of the conflict, in which no adverse effects are imposed on the transgressor, is available.[484] In terms of the framework developed above, anger can promote morally commendable outcomes with or without corresponding intentions. Apart from *actual* anger-motivated action, anger also generates norm conformity because potential perpetrators consciously or unconsciously (e.g., via conditioning) anticipate angry responses, which make anger-inducing actions more costly and can thus prevent them from ever happening.[485] To the extent that anger-motivated punishment has negative fitness consequences for the one being punished, it might help shape a population's gene pool towards less anger-inducing behavioral tendencies.[486] Angry responses on behalf of the self or others can motivate costly punishment of aggressors and free riders in collective tasks based on notions of reciprocity. Displays of anger alarm others of a problem and thereby facilitate collective action against transgressors; they can also express willingness to punish those who do not participate in first-order punishment.[487] The mere possibility of punishing free riders, even at a cost, can greatly increase cooperation. Free riding angers not only participants in the interaction, but also observers, and these onlookers can even be motivated to punish failures to cooperate. Such third-party punishment is common even in hunter-gatherer societies, while it does not seem to occur in nonhuman animals.[488] The approach motivation expressed in punishment and norm-restitutive behavior distinguishes anger from contempt and disgust, which tend to elicit avoidance.[489] Anger, in contrast to

---

483 Ibid., p. 126.

484 Prinz and Nichols quote a manuscript by Haidt and Sabini entitled "What exactly makes revenge sweet?"

485 See ibid., pp. 130–131.

486 See Boehm (2012), pp. 15–16.

487 See Jensen & Petersen (2011), p. 120. The last claim about second-order punishment is controversial. Boehm argues that classical hunter-gatherer communities do not punish those who fail to punish deviants, but that their systems of social control are nevertheless functional. See Boehm (2012), pp. 208–209.

488 See Prinz & Nichols (2010), pp. 130–131 and Jensen & Petersen (2011), p. 121. Punishment can be costly to the punisher for instance because of dangerous confrontations with the aggressor, or required use of resources. Note that *who* is being punished seems to depend on culture to some degree. While people from Western, industrialized countries and China punish free riders even at some cost to themselves and thereby foster cooperation, other populations engage in so-called *antisocial punishment* aimed at overly cooperative individuals, thus completely compensating the cooperation-inducing effects observed in Western populations. See Henrich et al. (2010), p. 70.

489 See Hutcherson & Gross (2011), p. 733.

disgust, can be attenuated by a sincere apology.[490] It relates to specific actions rather than individuals and it is generally less long lasting than disgust and contempt. This might be why people prefer being the object of anger to being the object of contempt, or worse, disgust.[491] Physiological measures support the hypothesis that anger is indeed separate from disgust: It produces a higher heart rate than disgust, also a much larger increase in finger temperature (which is not true for other heart-rate increasing emotions like fear or sadness).[492] The intensity of an anger response, in contrast to moral disgust, depends, among other factors, on the degree to which an individual perceives herself as affected by some behavior (the appraisal of self-relevance).[493] To the extent that the intensity of emotions affects the severity of moral evaluations, this might cause harsher judgments if the judge (or a close relative/friend/in-group member) is the victim of a transgression. High levels of arousal might also translate into impressions of 'seriousness' and lead to the classification of issues as moral rather than conventional. The mechanisms regulating third-party outrage appear to be attuned to the potential costs and benefits involved in an angry response, such as the risk of getting hurt in a confrontation, and the risk of being harmed in the future if one does *not* respond. Men get *angrier* with trustworthy exploiters in response to serious exploitative acts in the harm/fairness domain if the exploiter is more formidable. With respect to trivial matters however, formidability of the opponent *decrease*s anger.[494] There is some evidence indicating that more punishment is handed out the more 'outraged' subjects are,[495] though "kin, physically attractive, socially valuable and trustworthy individuals are treated more lenient [*sic*] when engaging in acts of exploitation."[496] Anger responses appear to be more sensitive to actual behaviors of others rather than to the underlying intentions, while disgust is tied to individuals who intended an outcome that is in violation of a moral code.[497] This makes evolutionary sense if punishment can alter future behavior of others. Even if the transgressor did not hurt an individual intentionally, an angry response will make him be more careful or consider the effects of his actions on others the next time around.

490 See ibid., p. 732.
491 See ibid., pp. 729–730.
492 See Schacter et al. (2009), p. 372.
493 See Suhler & Churchland (2011), p. 1211 and Hutcherson & Gross (2011), p. 726.
494 Formidability is the "capacity to inflict costs on others", see Jensen & Petersen (2011), p. 119. This is an interesting result because in nonhuman animals, opponent formidability increases the probability of deferring. One reason why humans might be different is their ability to form coalitions. See ibid., p. 120.
495 See Greene (2008b), pp. 52–53.
496 Jensen & Petersen (2011), p. 119.
497 See Hutcherson & Gross (2011), p. 720.

"While the punishment of accidents is guided by blindly retributive motives, it serves the farsighted function of modifying others' behavior."[498]

With respect to Haidt's criteria, anger is a prototypically moral emotion: It can be elicited on behalf of others, and its action tendencies can benefit those others and establish as well as stabilize a social order, both intentionally and unintentionally. In this sense, it certainly has an instrumental and most likely also a foundational relation to morality. The effect is foundational to the extent that appraisals that can make us angry on behalf of ourselves *and others* establish part of what is likely to be considered morally relevant. The *potential* other-relatedness serves to distinguish moral-foundational anger-eliciting appraisals from non-moral causes of anger. However, it is important to point out that many anger-eliciting appraisals are learned, and therefore not closely tied to adaptive problems in the EEA. I do not want to argue that *only* appraisals that occurred in the environment of our evolutionary ancestors can attain anger-based moral relevance. However, can we learn to feel moral outrage at the violation of a rule with just about any content? Even if that were possible in individual cases or for short periods of time, I suspect that it will be harder the farther psychologically removed the norm in question is from notions of physical or psychological harm, violations of reciprocity, or infringements of an agent's autonomy. This effect will be manifest in the notions of (anger-based) moral relevance that occur repeatedly across different times and cultures. In contrast, whether you, for instance, start wiping your desk on the left or on the right hand side will hardly ever count as morally relevant. Thus, the psychological processes involved in the experience of anger exert their influence on moral relevance by limiting the kinds of things to whose appraisal they respond, or, in other words, by determining the *aptitude* of things to appear morally relevant.

#### *4.4.1.2 Disgust*

Like anger, disgust counts as a basic emotion.[499] The evolutionary predecessors of disgust presumably lie in distaste reactions that prevent animals from contact with and ingestion of potentially harmful substances. Many disgust elicitors in modern humans still point to this evolutionary origin: North Americans, for instance, can be disgusted by "food, body products, animals, sexual behaviors, contact with death or corpses, violations of the exterior

[498] Cushman (2011), p. 262.
[499] See Rozin et al. (2008), p. 758.

envelope of the body (including gore and deformity), poor hygiene, interpersonal contamination (contact with unsavory human beings), and certain moral offenses"[500]. Some authors characterize disgust as central element of a 'behavioral immune system' evolved to minimize exposure to pathogens. Consequently, disgust can arise with respect to individuals with deviant physical appearance, people who engage in unusual behavior regarding food, sex, or hygiene, and members of out-groups.[501] This phenomenology could be the effect of two entangled motivational systems, one evolved to protect the gastrointestinal system from harmful food, the other, to protect the organism from contact with disease and parasites more generally. This evolved function might explain why the disgust system is more likely to produce false positives than potentially fatal false negatives.[502] Disgust motivates distancing oneself from the elicitor and is associated with nausea and a distinctive 'disgust face'.[503] Experiences of disgust in response to *moral* offenses might be a recent development that depends on the prior existence of a notion of morality. However, it is also conceivable that those, or at least some offenses we are disgusted with, became part of the moral domain *because* they elicited disgust, which would amount to a moral-foundational influence of this emotion. Inbar and Pizarro distinguish three hypotheses regarding the relation between disgust and morality: Firstly, disgust might arises *as a consequence* of the perception of a moral violation. Secondly, disgust might function as an *amplifier* of extant moral condemnation. Thirdly, in some cases disgust could actually *cause* moral condemnation. These relations are not mutually exclusive; several might be true.[504]

An influential hypothesis about the phylogenetic development of disgust is based on a notion of *core disgust*, an emotion that originates from a distaste-based rejection mechanism for food. Core disgust is elicited by simultaneous appraisals of potential oral incorporation, offensiveness, and danger of contamination.[505] Neurophysiological evidence supports the idea that core disgust in animals and humans and social disgust in humans are related: The human insular cortex, the region most frequently associated with food- and nonfood-related disgust (it is also associated with anger and autonomic arousal), probably evolved from the

---

500 Ibid., p. 757. See also Haidt (2003a), p. 281, drawing on Rozin: "[…] disgust is best understood as a complex emotion that protects the body and the soul from degradation." It is worth noting that while vomit and feces are near-universal disgust elicitors in adult humans; neither infants nor nonhuman animals regularly avoid them; they can even be attracted by them. The fact that these disgust-responses are not present from birth could mean either that they are learned, or processed by later-developing psychological modules. See Rozin et al. (2008), p. 765.

501 See Inbar & Pizarro (2014), p. 115.

502 See Kelly (2014), p. 134.

503 See Rozin et al. (2008), pp. 758–759.

504 See Inbar & Pizarro (2014), p. 112.

505 See Rozin et al. (2008), p. 759. Contact with the disgusting object renders food unacceptable. Primary offensive objects are all animals and their products. See ibid., pp. 757–760.

so-called gustatory cortex in primates, which has a role in the selection of food.[506] In humans, the insular cortex has been associated with moral issues such as the evaluation of sexual behavior. Insular activation also increases in response to unfair offers in ultimatum games, correlates with percentage of rejected offers, and is attenuated when subjects believe that a computer rather than a real person made the offer.[507] Apart from the insula, the experience of disgust has been linked to activity in the basal ganglia and some areas of the prefrontal cortex.[508] Physiological evidence indicates that anger and disgust are distinct emotions: Disgust produces stronger galvanic skin response (more sweating) than anger,[509] but in contrast to anger and fear, it lowers the heart rate (association with parasympathetic responses).[510]

The class of disgust elicitors in humans extends far beyond phenomena associated with core disgust such as food, body products, and animals.[511] As a hypothesis about how human beings came to be disgusted with so many things, Rozin et al. proposed the concept of animal-nature disgust[512]: Eating, excreting, sexuality, and death are highly regulated in most cultures. Violations of the corresponding rules are met with disgust, and all these behaviors supposedly remind us of our animal nature. In contrast to core disgust, animal-nature disgust does not require potential *oral* incorporation, but rather physical contact more generally. However, Rozin et al. do not provide much of an explanation for why we should have come to detest our animal nature, apart from speculation about "our fear of animal mortality"[513].

While less basic or evolutionarily ancient than those of core disgust, the elicitors of animal-nature disgust are still related to the body. However, it seems that people can feel disgust also towards, for instance, betrayal or racism, which are not as easily linked to animal nature. Is there such a thing as specifically *moral* disgust? Some have argued that the lay use of disgust is close to the theoretical meaning of *core disgust* when elicitors are related to the body, but closer to the theoretical meaning of *anger* when elicitors are social events without salient relation to the body.[514] If that were true, moral anger might be a more appropriate label for the emotion people experience at least in response to violations of moral norms

506 See ibid., p. 758.
507 See Greene (2008b), p. 54.
508 See Rozin et al. (2008), p. 768. The basal ganglia are a group of nuclei at the base of the forebrain associated with a variety of functions, including voluntary movement and emotion.
509 See Schacter et al. (2009), p. 372.
510 See Rozin et al. (2008), pp. 758–759.
511 See ibid., p. 764.
512 See ibid., pp. 758–759.
513 See ibid., p. 764.
514 See Royzman et al. (2009), p. 166.

which are not related to the body, but for instance to notions of fairness. Rozin et al. quote insular activity in response to unfair ultimatum-game offers as indicative of moral *disgust*. However, as mentioned above, this area has also been associated with anger; therefore these findings are inconclusive.[515] The fact that the respective words for disgust are applied to social behavior in many languages provides more promising evidence for a specific role of disgust in moral evaluations.[516] Other physiological responses are also informative: Subjects who claimed to be more disgusted than angered by a video about American neo-Nazis showed a decrease in heartrate, rather than the increase typical of anger.[517]

Cultural processes have greatly enlarged the class of disgust elicitors; triggers of moral disgust in particular can be very different from activators of health-related core disgust, although there is also overlap. Norms that regulate what it is proper to ingest contingent on social roles, gender, or season play an important role in tribal and religious codes of conduct. Humans can learn to experience disgust towards certain kinds of meat, physical contact with members of particular social classes, specific sexual practices, etc. How these disgust-eliciting violations are construed depends, among other factors, on the prevalence of religion in public awareness: While disgust arises as response to violations of deity-imposed norms that demand purity in religious societies, secular societies often construe violations of purity norms as "crimes against nature"[518]. Although the variety of cultural concepts that involve disgust appears very wide, many of them share a concern about what body and mind come into 'contact' with. Sometimes, (moral) evaluations mirror the language used to refer to physical elicitors of disgust (e.g., 'dirty lies'). According to the CAD-triad hypothesis, disgust is elicited by transgressions against norms that belong to Shweder's ethic of divinity. In Haidt and Joseph's more recent framework, disgust is most closely associated with the norms of sanctity/degradation, but it can also be elicited by other transgressions, like betrayal and treason (loyalty/betrayal and authority/subversion) or cruelty (care/harm).[519] Transgressions against persons elicit disgust in case they are particularly violent, unmotivated, or spring from 'demeaning' motives.[520] Inbar and Pizarro state that so

[515] See Rozin et al. (2008), pp. 762–763.
[516] See ibid.
[517] See ibid., p. 763.
[518] Prinz & Nichols (2010), p. 122.
[519] Hutcherson & Gross (2011) present evidence indicating that disgust might be an adaptive response to moral violations generally, rather than only to violations of norms that belong to the domain of sanctity/degradation. However, this might be an artificial effect since they contrasted '*moral* disgust' with 'anger' and 'contempt' (sans '*moral*'). See ibid., p. 724.
[520] See Prinz (2007a), p. 178.

far empirical evidence is in line with the notion of disgust as the "guardian of physical purity", but does not support the thesis that disgust arises also in response to violations of spiritual purity.[521] In its function as protector from noxious substances, disgust encompasses a "motivation to avoid, expel, or otherwise break off contact with the offending entity, often coupled to a motivation to wash, purify, or otherwise remove residues of any physical contact that was made with the entity"[522]. An experiment that found subjects uncomfortable with the idea of wearing a laundered sweater that had allegedly been worn by Adolf Hitler strikingly illustrated the transferal of such behavior to moral offences.[523] Apart from immoral behavior, other properties of the previous owner, such as disease, misfortune, or strangeness also make laundered sweaters less desirable (interpersonal disgust).[524] These examples also show that disgusting entities are contagious. From about the age of three to five years, humans become sensitive to the contact history of objects: We avoid contact with things that have been in contact with original disgust elicitors, and beliefs about disinfection have little effect on that tendency.[525]

Moral disgust can be 'prosocial' in the sense of stabilizing social order, for instance when it motivates punishment of socially damaging behavior through ostracism, or as part of an emotional punishment-and-reward structure that stabilizes separation between social groups. Whether these instrumental effects depend on awareness of a moral code might depend on how similar the disgust elicitor in question is to the original triggers of core- and animal-nature disgust. The presence of disgust towards a person predicts ascriptions of immoral *character* better than both anger and contempt, and disgust-based evaluations appear to be less sensitive to apologies or attempts to make amends than those grounded in anger.[526] More so than contempt and anger, disgust is a response to *intentional*, immoral behavior. Contempt, in contrast, tends to be elicited by displays of incompetence.[527] More so than anger, which promotes costly, active responses to immediately threatening and harmful behavior of others, disgust and contempt motivate the less costly behavior of avoiding risky contact with those who have displayed pernicious behavior in the past.[528] Compared with other emotions, the actual experience of disgust seems to be rather short-lived.[529] However,

521 See Inbar & Pizarro (2014), p. 121.
522 Haidt (2003c), p. 857.
523 See Jones (2007), p. 769.
524 See Rozin et al. (2008), p. 762.
525 See ibid., p. 560, Jones (2007), pp. 768–769.
526 See Hutcherson & Gross (2011), p. 732.
527 See ibid., p. 733.
528 See ibid., p. 720. Ibid. take this to be the primary function of contempt and disgust.
529 See Rozin et al. (2008), p. 759.

Hutcherson and Gross suggest that social/moral disgust may generate more long-lasting judgments about persons than episodes of anger do, because disgust is a response to rather *stable* intentions or character traits, while anger is elicited by *immediate consequences* of an action.[530] In this respect, disgust resembles shame, while anger corresponds to guilt (see section 4.5.1).

There is some evidence that disgust is frequently involved when new vices are identified.[531] Rozin et al. suspect that, since links between elicitors of core- and animal-nature-disgust and those that evoke moral disgust are often far from obvious, the role of disgust in these matters might be due to an episode of 'preadaptation', a process in which an extant adaptation acquires a new (additional) role. For instance, teeth and tongue are adaptations for food consumption. More recently, they have come to play a major role in speech. Reference to 'preadaptations' enables Rozin and his coauthors to put a label on the extension of the class of disgust elicitors both in cultural evolution and ontogenesis.[532] It does not, however, explain why disgust is *more* likely to figure in such moral preadaptations than other emotions (if it is).

Evidence indicates that disgust influences moral judgments negatively: In the study by Wheatley and Haidt mentioned in section 2.2, hypnotically induced disgust made negative moral judgments more severe or even *generated* negative evaluations that were absent without that disgust. Haidt et al. observed that disgust is an emotion frequently elicited by the food- and sexuality-related harmless taboo violations subjects evaluated in their studies (eating a chicken used for masturbation or dead pets, etc.). At present, it is unclear whether disgust selectively amplifies moral condemnation concerning *specific* subject matters, or whether the effect is general.[533] However, North American subjects of high socioeconomic status (SES) separated their judgments from this emotion, while disgust corresponded with negative moral evaluation in all other groups tested (high and low SES in Brazil, low SES in North America). Thus, the influence of disgust on moral judgments appears to be mediated by culture.[534] In particular, Haidt and his coauthors hypothesize that disgust elicitation alone does not make an action wrong in cultures with harm-based moralities, since in these cases moral transgressions require a victim.[535] Disgust also plays a role in the relations between different groups: It is often felt towards groups that are perceived as being lower in

530 See Hutcherson & Gross (2011), p. 728.
531 See Rozin et al. (2008), p. 763.
532 See ibid., p. 764.
533 See Inbar & Pizarro (2014), p. 122.
534 See also Rozin et al. (2008), p. 766.
535 See Haidt et al. (1993), p. 21.

status and generally unlike the in-group. For instance, people feel disgust towards foreigners or those perceived as sexually deviant.[536] Conversely, disgust is increasingly suspended the more intimate relationships become. People care for their sickly relatives; parents regularly get in contact with their infant offspring's excrement. While the 'objects' remain disgusting and stronger motives override the aversive response in these particular scenarios, otherwise disgusting events can actually be perceived as pleasant in sexual relations.[537] Individuals and groups differ in their disgust sensitivity: Women are more sensitive than men, young adults are more sensitive than old adults; higher socioeconomic status, better education, and greater openness to experience are negatively correlated with it, while it increases with neuroticism. In line with the behavioral-immune-system hypothesis, people who feel more vulnerable to disease and who live in areas with higher historic disease prevalence tend to have more conservative sociopolitical views, potentially mediated by increased disgust sensitivity.[538]

Does disgust motivate behavior that produces morally commendable outcomes? I believe it does. It directly motivates us not to engage in activities we have been hardwired or educated to consider disgusting. Some of them, like incest, are considered morally wrong. Indirectly, the anticipation of disgusted responses of others will have similar effects, although this effect need not co-occur with an experience of disgust in the agent, only with a motivation to avoid the kind of behavior towards the agent that would result from the others' disgust experience. As for foundational influences of disgust on the moral domain, I find it quite plausible that this influence is manifest in moral prohibitions of actions that involve the kinds of things that were elicitors of the phylogenetically early variants of disgust, such as norms that involve food, sexuality, or the treatment of corpses. On the other hand, it seems that the capacity for disgust can be recruited to establish new moral concerns whose relation to those evolutionarily shaped elicitors is not similarly straightforward.[539] One might argue, however, that these new concerns can only gain behavioral traction if they are successful in recruiting the motivational resources which disgustability provides. Framing candidate issues in terms that create a connection to more ancient elicitors of disgust (rotten, dirty, slimy, foul, etc.) could have such an effect.

536 See Rozin et al. (2008), p. 770.
537 See Ekman (2003), pp. 177–180, also the footnote on p. 178.
538 See Inbar & Pizarro (2014), p. 117.
539 Kelly (2014), p. 134 claims that disgust is particularly malleable and responsive to social influence.

#### *4.4.1.3 Contempt*

Contempt is sometimes characterized as a blend of anger and disgust that involves feeling superior and looking down on others.[540] In social contexts, it typically is a response to conduct judged inappropriate with respect to social roles and hierarchies. Its evolutionary roots might lie in the motivations that enable reciprocal altruism in animals: People feel contempt for others who do not live up to the standards of conduct befitting them; objects of contempt are met with 'cool indifference', their affairs are considered unworthy of attention and excitement. Whereas anger and disgust are associated with high levels of arousal, contempt brings less arousal and acts as social-cognitive abatement of positive other-conscious emotions.[541] We feel less compassion, warmth, and respect for those of whom we are contemptuous.[542] In contrast to disgust, contempt is not necessarily an unpleasant experience.[543] Those who hold that contempt is a blend of anger and disgust argue that the elicitors of contempt combine elicitors of these emotions: Violations of role-dependent norms are hybrids. They constitute both a transgression against the natural order manifest in social hierarchies (crime against nature, typical elicitor of disgust) as well as an offense to others (a crime against an individual, typical elicitor of anger).[544] Others who characterize contempt as an independent emotion, such as the proponents of the CAD-triad hypothesis, discern differences between the elicitors of anger, disgust, and contempt. While contempt (also guilt and shame) tends to be a response to transgressions that harm the community, anger is more closely related to 'crimes against individuals'; violations of the natural/sacred order are met with disgust.[545] These authors have therefore hypothesized that in contrast to anger, the adaptive effects of the expression of disgust and contempt mainly work not by motivating the agent whose behavior elicited the emotion to change his ways, but importantly also through third parties: Displays of anger seem particularly apt to stop the anger-inducing behavior. Displays of contempt, in contrast, might primarily serve as signals in a reputation system, because they relate to violations that are less contingent on (short-term) individual interest. We avoid interaction with the disgusting and those upon which we look with contempt.[546] There is some evidence that contempt is generally a response to displays of incompetence, and that contempt for 'immoral' behavior that fails to observe community

[540] See Prinz (2012), p. 306.
[541] See Haidt (2003c), p. 858.
[542] See ibid.
[543] See Ekman (2003), p. 182.
[544] See Prinz (2012), pp. 306–307.
[545] See Prinz (2007a), p. 178.
[546] See Hutcherson & Gross (2011), p. 734.

standards related to status etc. is just a subclass of these incompetence-related responses.[547] Even if anger, contempt, and disgust differ in their characteristics, they nevertheless often arise in parallel; they appear to be quite similar when compared to other emotions.[548] Neurological findings are compatible with this suggested similarity. As mentioned above, activity in the insular cortex is associated with both anger and disgust. The dorsolateral prefrontal cortex and the lateral orbitofrontal/perirhinal cortex are associated with both anger and contempt.[549]

How does contempt relate to morality? I believe it is foundational insofar as its coupling to specific elicitors establishes a concern with norms that govern hierarchies and social roles. These concerns can, and presumably did, arise without awareness of a moral code. Once they establish part of a moral code, however, awareness of its requirements can probably also spur contempt caused by impressions of incompetence with respect to the moral code (a kind of rule-related extra-contempt). I reckon that the foundational influence rests on intrinsic motivation, since the behaviors brought forward by contempt (disregard for the object of contempt, gossip, etc.) do not appear to be motivated by considerations of external aims. As with disgust and anger, the anticipation or awareness of these other-critical responses and the respective behavioral consequences in others resulting from one's actions can motivate agents to adjust their behavior.

#### *4.4.1.4 Jealousy*

By Haidt's criteria (disinterested elicitor, prosocial action tendency), jealousy is not a moral emotion: It is seldom felt on behalf of others. Nevertheless, it strongly shapes norms of interaction that count as moral. It is elicited by others perceived as a threat to important social relationships of the self and involves aggression aimed at removing that threat. Animal aggression against sexual rivals appears to be an obvious behavioral analogue. The adaptive function of such an emotion is rather straightforward: For males, aggression against sexual rivals serves to ensure access to the female, and to rule out that investment in care for offspring actually helps a competitor's propagation. For females in pair-bonding

[547] See ibid., p. 732. However, "[a]lthough contempt was clearly linked to incompetence […], it may be that this is only one of a number of necessary eliciting appraisals for it. Simply being incompetent may be enough to elicit sadness, pity, or amusement […], but to elicit contempt may require something more, including but not limited to a judgment of moral laxness, an unsympathetic nature, or a competitive relationship to the perceiver […]." Ibid., p. 733.

[548] See ibid., pp. 733–734.

[549] See Moll et al. (2008a), p. 168. The perirhinal cortex is a part of the medial temporal lobe.

species, securing a mate that provides protection and raises offspring is similarly essential.[550] The experience of jealousy is a mix of hurt, anxiety, and anger, combined with an appraisal of betrayal.[551] In effect, jealousy can stabilize relationships, both because of the actual behavior it triggers on the part of the jealous agent and because rivals and partners anticipate such aggressive reactions. Moral norms regarding obligations in romantic and sexual relationships between men and women in particular owe their shape to the mechanisms of jealousy to a significant degree. In terms of Haidt's foundations, these norms relate to notions of reciprocity and fairness. As far as I am aware, however, Haidt and Joseph's publications on moral foundations theory do not mention jealousy. Valdesolo and DeSteno classify jealousy, as well as pride and vengeance, as moral emotions since they enhance general well-being by generating the social cohesion required for "flourishing cooperative societies"[552], even though at first glance they only promote self-interest. These effects, however, point only to an instrumentally moral character of jealousy, since the morally commendable outcome (general flourishing) is unintended. Let us disentangle these effects: Evolutionary processes tied jealousy to specific elicitors. These eliciting events appear morally relevant because they trigger strong emotional responses. This is the foundational influence. Acts motivated by jealousy and the anticipation of such acts may serve to stabilize social relationships; the aggressive behavior motivated by it "serves as an honest signal to partners of a strong degree of psychological investment in a relationship."[553] In this sense, jealousy can direct behavior to conform to the requirements of fidelity established by the emotion itself. Valdesolo and DeSteno's argument offers a different justification for calling jealousy a moral emotion: If it stabilizes social relations, and social stability is considered morally commendable for reasons not directly related to jealousy (for instance because it is conducive to general well-being), then jealousy also has a morality-promoting effect through outcomes that are unintended in jealousy-motivated actions.

[550] These differences in parental investment have led to the hypothesis that women and men differ in their concern for infidelity: Men tend to worry more about sexual infidelity, while women care primarily about their partner's emotional attachment to them. This fits with the observation that men kill women much more frequently than vice versa, and that actual or impending rejection by a sexual partner and infidelity are the most common motives for murder. See Ekman (2003), pp. 130–131.

[551] See DeSteno et al. (2006), p. 627.

[552] Valdesolo & DeSteno (2011), p. 276.

[553] Ibid., p. 277.

### 4.4.2 Other-Praising Emotions: Gratitude and Elevation

*Other-praising* emotions like gratitude and elevation can be elicited by moral behavior and motivate moral behavior. Gratitude, in particular, involves an appraisal of benevolent behavior of others towards the self or close relations which is often intensified when the "benefits are unexpected and/or costly to the benefactor"[554].[555] Contrary to indebtedness, it is pleasant.[556] The evolution of gratitude, like contempt, is probably connected to the emergence of reciprocal altruism: It motivates agents to repay benefits for themselves or those they feel close to and can thereby reinforce valued behavior in the benefactor; however, the prosocial motivation can also extend beyond the benefactor. Some researchers distinguish between generalized and benefit-triggered gratitude: While the latter arises in response to specific benevolent acts, the former is a reaction to what is more generally considered valuable in one's life.[557] Benefit-triggered gratitude increases the willingness to engage in helping behavior.[558] In economic decisions in which pit personal gains (not the status quo, as in requests for assistance) against personal losses, such gratitude increases cooperative behavior. This effect appears to go beyond mere adherence to a reciprocity norm or strategic considerations, since the increase in cooperativity occurs in one-time interactions with strangers as well as in interactions with the benefactor.[559] Mere positivity or happiness does not consistently have that effect.[560] From an evolutionary point of view, the increase in cooperativity towards strangers could be a by-product of the adaptive effect that concerns cooperation with the benefactor.[561] Generalized gratitude among partners in close relationships (e.g., marriage) that results from the perception of costly relationship maintenance behavior increases such maintenance behavior in return.[562] This is interesting because partners in close relationships do not usually keep track of costs and benefits of the relation as keenly as people often do in exchanges with strangers.[563] The effect of benevolent acts on the recipient seems to be less significant than the effect of interactions perceived as negative. We recall bad events more easily, perception of others is more strongly affected by their undesirable rather than their praiseworthy behavior, and immoral behavior is more

554 Tangney et al. (2007), p. 362.
555 See Kubacka et al. (2011), p. 1363.
556 See Tangney et al. (2007), p. 362.
557 See Kubacka et al. (2011), p. 1363.
558 See DeSteno et al. (2010), p. 289.
559 See ibid., p. 291.
560 See ibid., p. 290.
561 See ibid., pp. 292–293.
562 See Kubacka et al. (2011), p. 1371.
563 See ibid., p. 1363.

easily attributed to the agent herself, while conformity with moral requirements is often seen as due to situational pressures.[564]

Jonathan Haidt also mentions elevation as a moral emotion. Elevation arises when we witness moral excellence in acts of remarkable charity, loyalty, or self-sacrifice but we ourselves are not necessarily beneficiaries of that act. It creates a desire to excel.[565] While experiments have confirmed the prosocial action tendencies resulting from elevation experimentally, it has not yet been determined how long that motivation lasts.[566] Haidt characterizes it as the opposite of social disgust with respect to both elicitors and action tendencies: It is triggered by acts that surpass expectations, while disgust is elicited when people fall short of these standards; it motivates to seek contact, while disgust makes people move away from the elicitor.[567] Paul Ekman agrees that such a phenomenon exists, but is not certain that it qualifies as a proper emotion. The sensibility for elevation might differ across individuals: People for whom adherence to moral standards is an important part of their self-concept (a quality referred to as moral identity) perceive helping behavior in others more positively.[568] Jesse Prinz suggests that positive emotions can be characterized with reference to *agent* and *patient* of a good deed. Instead of elevation, he mentions admiration as a second positive emotion that arises in response to morally praiseworthy behavior. While gratitude is elicited if somebody else benefits the self, admiration occurs when we observe somebody benefitting somebody else. A good deed done by the self to somebody else elicits gratification.[569] Haidt's elevation is supposed to arise in response to actions in which somebody else benefits from actions, either one's own or somebody else's.

### *4.4.3 Other-Suffering Emotions: Empathy and Sympathy*

Haidt's conception of *other-suffering* emotions comprises concepts like sympathy, empathy, or compassion that were emphasized by the sentimentalists of the Scottish enlightenment such as David Hume or Adam Smith, as well as by cognitive-developmental moral psychologists like Piaget and Kohlberg.[570] Since 'empathy' in particular notoriously denotes many different things in psychology, I will first characterize some important concepts associated with the term and then discuss their respective role with regard to morality.

564 See Aquino et al. (2011), p. 703.
565 See Haidt (2003c), p. 864, Tangney et al. (2007), pp. 361–362, and Aquino et al. (2011), p. 715.
566 See ibid., pp. 715–716.
567 See Haidt (2003c), p. 864.
568 See Aquino et al. (2011), p. 704.
569 See Prinz (2007b), pp. 81–82.
570 See Haidt (2003c), p. 861.

The origins of empathy and sympathy, or rather, this complex of emotional capacities, presumably lie in the attachment system which makes animals sensitive for the well-being of their kin and motivates them to eliminate causes of suffering or offer consolation. In humans, cooperative breeding has supposedly led to particularly pronounced empathic capabilities.[571] All of these empathy-related notions comprise a response to the state of another in an observer. Various accounts, however, address different types of responses. A few characteristics of these accounts are particularly distinctive: Firstly, they require *congruence* between the emotional state of the target and the observer to different degrees. The spectrum ranges from identical emotional experience, experience in the observer that is merely similar in valence, to phenomena that do not require congruence with the actual emotions of the target, but rather with the emotions she would experience were she aware of her situation in the way the observer is.[572] Secondly, notions of empathy differ with respect to the kind of information processing required. While some empathic phenomena occur automatically without conscious deliberation, others require that effort be put in thinking about and understanding the target's predicament.[573] A specific aspect of the differences in cognitive requirements is the extent to which the various empathic capacities require awareness of the difference between the self and the other, i.e., awareness of the observer regarding the source of his experience.[574] The cognitive demands associated with different empathic capacities affect the extent to which similar psychological capacities can be found in related and ancestral species, which is relevant in turn for the evaluation of the evolutionary explicability of morality. Thirdly, empathy-related phenomena differ in their orientation: while some go along with reactions directed primarily towards the self, others appear to be more other-oriented and correlate, for instance, with helping behavior.

A basic phenomenon is *emotional contagion*.[575] Here, congruence between the emotional state of the observer and the target is very high. Emotional contagion is typically thought to require direct perception of some sensory input that expresses the target's emotional state. An oft-mentioned example is that infants in maternity wards respond with crying to the crying of other infants.[576] This kind of response is frequently characterized as empathetic reaction that does *not* require a distinction between the self and the other: The infant need not be aware that utterances of another individual are the *source* of her distress, it also need

[571] See Voland & Voland (2014), p. 119. Cooperative breeding refers to nonparents caring for offspring.
[572] See Stich et al. (2010), p. 171, Preston & De Waal (2001), p. 4.
[573] See ibid., p. 3.
[574] See Stueber (2008).
[575] See ibid.
[576] See Preston & De Waal (2001), p. 7, Kitcher (2011), p. 28.

not be aware of the *situation* of the other infant. The finding that the response of infants to signs of distress in others changes with time corroborates this interpretation: While at a very young age, infants respond to signs of distress by becoming distressed themselves and seeking consolation, they start to attend to the distressed individual in the second year of their lives.[577] This onset of helping targeted at the needs of the victim has been linked with the occurrence of mirror self-recognition (MSR), which indicates self-awareness and an element of self-other differentiation.[578] However, in other experiments, one-day-old babies responded most strongly to the crying of other newborns, while they responded less strongly to the crying of an 11-month old, and not at all to recordings of their own crying, which might indicate the presence of a rudimentary self-other distinction at birth.[579] This ability to respond with similar emotions to the emotional experience of others is a basic component of *affective* notions of empathy. The responses observed in infants display one manner in which these reactions can come about, namely through direct perception of emotion cues in others. I address other, more cognitively demanding ways of generating congruent emotional responses in the next paragraph. It has been suggested that this rather basic emotional contagion involves mirror neurons: These neurons 'fire' both when executing a particular kind of behavior, *and* when it is merely *observed*. This explanation of emotional contagion fits well with emotion viewed in the spirit of the James-Lange tradition: Advocates of the so-called facial-feedback hypothesis suspect that the distinctive 'feel' of a particular emotion just *is* the experience of the physiological changes that constitute this emotion, including the corresponding facial expression. They claim that empathy works at least partly via subconscious mimicking of the other individual's facial expression, which generates a corresponding subjective experience. Support for this view comes from findings that individuals high in empathy are also more susceptible to yawn contagion, while such contagion is absent in children with empathy deficits.[580] This pattern of explanation also extends to posture and body language.[581]

Other variants of empathy depend less on the direct perception of others, but require more highly developed cognitive capacities. *Perspective taking* refers to the ability to both

577 See Haidt (2003c), p. 861.

578 See De Waal (2008), pp. 285–286, Bierhoff (2002), pp. 119–122. The standard test of MSR requires the subject to realize that a mark painted on it's cheek, which it can see only in the mirror, is actually on it's own cheek (indicated by the attempt to remove it from one's *own* face), rather than on the face in the mirror. MSR has been observed in great apes, dolphins, and elephants. See De Waal (2008), p. 286, De Waal (2013), pp. 115–116.

579 See Decety & Jackson (2004), p. 78.

580 See De Waal (2013), p. 138.

581 See Schacter et al. (2009), p. 381, Preston & De Waal (2001), pp. 11–12.

imagine the situation of others and attribute mental states to them in order to evaluate the imagined situation based on beliefs, intentions, and desires attributed to the target.[582] Some authors use the term *cognitive empathy* to refer to this ability, where *empathic accuracy* indicates the extent to which an individual is able to identify the mental states of the target.[583] Perspective taking does not necessarily involve congruent emotional arousal in the observer. Whether congruence occurs presumably depends on the psychological mechanisms employed to understand the other's situation. If the understanding is based on, say, direct perception of emotion cues, accounts of emotional cognition that emphasize physiological and experiential aspects might expect a similar emotional experience in the observer. However, this response can apparently be suppressed or overridden: Think of hateful torturers or dutiful physicians applying painful procedures. Moreover, constructing an image of another's situation from less vivid input such as a dry verbal description in a book might be less likely to generate congruent emotional responses. In any case, perspective taking is an essential requirement for many instances of *affective empathy*.[584] Affective empathy involves an emotional response in the observer that is appropriate to the perceived situation of the target and caused by an appreciation of the target's situation.[585] In case both the target and the observer have a similar understanding of the situation the target is in, this implies congruence between the emotional experience of the target and the observer. In case the observer has a different understanding of the situation of the target than the target itself, this kind of congruence need not occur, for instance if the target is still blissfully unaware of a disaster that the observer can already apprehend. It is important, however, that even if emotions between observer and target are not congruent, the emotional response of the observer is based on the mental states attributed to the target rather than his own (what would the target feel if she believed what the observer believes?). Unlike emotional contagion, affective empathy requires that the self and the other remain distinct in the mind of the observer (as perspective taking does).[586] The observer is aware that the situation of the other

582 See De Waal (2008), p. 285.
583 See Bierhoff (2002), pp. 135–138.
584 See Stueber (2008).
585 A well-known definition of affective empathy by Hoffman (quoted in ibid.) involves a comparative component: Empathy involves "[…] feelings that are more congruent with another's situation than with […] [my] own situation." This definition is unable to capture affective empathy with others that are *in the same situation* as the observer. In order to avoid these difficulties, I avoid the comparative element and refer to the *cause* of the observer's emotion.
586 See ibid.

is the source of his experience. This does not imply, however, that the psychological mechanisms that enable emotional contagion cannot contribute to the affective component of affective empathy.

Recent publications often contrast empathy (used in senses similar to affective empathy) with *sympathy*.[587] Like (affective) empathy, sympathy involves a clear distinction between the self and the other. Karsten Stueber argues that in contrast with affective empathy, sympathy does *not* require a congruent emotional experience. Rather, it is "an emotion *sui generis* that has the other's negative emotion or situation as its object from the perspective of somebody who cares for the other person's well being [*sic*]."[588] Prinz defines sympathy as a "negative emotional response to the suffering of others."[589] Does this suffice to distinguish affective empathy and sympathy? As we have seen, affective empathy equally does not require an emotional experience that is congruent with the target's *actual* experience, but 'merely' an emotional response based on the observer's understanding of the target's situation as per the mental states attributed to the target. According to this definition, it is hard to keep the two phenomena apart if the emotional response of the observer is congruent with either the actual or the imagined response of the target. Clearer examples of sympathy, as distinct from affective empathy, occur when concern with the other's well-being is expressed without pronounced affective reaction, or if the response is not accompanied by an affect of the same valence as the target's (actual or simulated) emotion (for instance if a good-humored individual attempts to cheer up his gloomy companion without becoming depressed himself). Moreover, affective empathy seems to require a response that is more tightly connected to the actual preferences of the target in a given situation. In contrast, sympathy allows for a larger degree of paternalism. For instance, a sympathetic parent who is sincerely concerned about the well-being of his child bases his actions on the desires and beliefs he *would like* the offspring to have. Responses to unconscious accident victims illustrate another difference between sympathy and empathy. Since unconscious individuals have no present emotional state, they can be subjects of sympathy, but not of empathy.[590] However, on Stueber's liberal definition of affective empathy, concern for unconscious individuals could count as instance of empathy, since he permits that the emotional response of the observer is incongruent with the actual response of the target. The observer's characteristics are also

[587] See Bierhoff (2002), pp. 107–108. This reverses a trend in psychology to discuss only empathy, which led to increasing conceptual confusion. See ibid.
[588] Stueber (2008). Emphasis in the original.
[589] Prinz (2007b), p. 82.
[590] See Stich et al. (2010), p. 171.

relevant: Individuals incapable of certain emotions, or of experiencing pain, cannot empathize with others who experience these sensations, but they can sympathize.[591] In addition, it seems possible to sympathize with an angry person without becoming angry oneself.[592] Definitions of sympathy frequently refer only to negative emotions or distress of the target. I believe it makes sense to conceive of sympathy as a response that also relates to *positive* affect in the target. If I am concerned with another's well-being, not only his despair, but also his happiness is relevant.[593]

Let me mention one more concept that might help clarify the contours of sympathy: *Personal distress* is a negative emotional response to the negative emotion or situation of others.[594] It differs from sympathy in terms of action tendencies: While sympathy supposedly motivates acts targeted at improving the well-being of the other, personal distress motivates whatever action constitutes the 'cheapest' way to end the unpleasant experience caused by the other's distress. This can lead to helping behavior, but it can also make the observer leave or distract himself from the other's plight.[595] Even if the personally distressed individual is aware that somebody else's problems *cause* her distress, this distress becomes her own. The empathic individual, in contrast, recognizes that she is not herself in distress. The personally distressed individual is more concerned with her own distress than with the distress of the target.[596]

The phenomena discussed are not mutually exclusive. The dimensions of distinction (self-other differentiation, congruence of experience, cognitive mechanisms involved, self- or other-orientation) can occur in various combinations, and can be realized to different extents. Moreover, according to the terminology employed here, both sympathy and empathy are not emotions strictly speaking, since they lack specific valence (particularly if sympathy refers to reactions to positive as well as negative affect) as well as specific action tendencies. One might add that elicitors are also quite diverse. It seems more adequate to conceive of them as capacities or dispositions to experience certain emotions in response to specific (actual or imagined) emotional experiences of others.[597]

---

591 See Prinz (2007b), p. 83.
592 See Tangney et al. (2007), p. 363.
593 See Stich et al. (2010), p. 172.
594 Haidt (2003c), p. 862 mentions a very similar phenomenon called "distress at another's distress".
595 See Stich et al. (2010), p. 171, Stueber (2008), De Waal (2008), p. 283, Turiel (2006a), p. 800. While Stueber contrasts personal distress with empathy since it renders Hoffman's notion of greater appropriateness to the situation of the other pointless (the personally distressed individual is herself distressed), de Waal describes personal distress as "born from empathy".
596 See Tangney et al. (2007), p. 363.
597 See ibid., p. 362.

Many authors have attributed vital roles in morality to other-suffering emotions or capacities. For instance, psychologist Daniel Batson proposed a so-called empathy-altruism hypothesis.[598] Batson discusses empathy as an aversive emotional response to the suffering of others that goes along with concern for their well-being (other-directed). In experiments, it is elicited by either emphasizing similarity between the target and the observer or instructing the observer to imagine what the target feels.[599] The empathy-altruism hypothesis claims that whether we help people in need even if it is not in our self-interest to do so (cost outweighs benefits to the self) depends on the presence of empathy for the person in need. Batson's research on this topic is embedded in a debate about the existence of psychological altruism. He attempts to determine whether helping behavior is ultimately the product of altruistic desires for the well-being of others or merely a means to increase our own well-being (for instance if we help just because being confronted with suffering is unpleasant, because we desire praise for our nobility, etc.). While this project is not yet finished, it has weakened the plausibility of two popular egoistic explanations of helping behavior: the aversive-arousal-reduction hypothesis (when we feel empathy, we help in order to reduce our own aversive experience) and the socially-administered-empathy-specific-punishment hypothesis (when we feel empathy, we help in order to avoid punishment for not helping by our community).[600] Within this debate, it is uncontroversial that helping behavior is more frequent when we feel empathy, as induced by making subjects read about the personal values of the victims, imagine being in their situation (perspective-taking), and emphasizing similarity between subject and victim.[601]

Psychologist James Blair also believes that empathy is quite important for the development of our moral capacities. His theory is founded on research on psychopathy, "[…] a developmental disorder that involves emotional dysfunction, characterized by reduced guilt, empathy and attachment to significant others, and antisocial behavior including impulsivity and poor behavioral control."[602] Psychopaths tend to be irritable and aggressive, callous, and do not appear to feel remorse. Interestingly, they also appear to be less able to distinguish between moral and conventional rules like 'normal' members of Western populations can. Instead, they often treat *all* rule transgressions as serious and prohibited independent

598 See Stich et al. (2010) for a lucid discussion of Batson's research.
599 See ibid., pp. 172–181, Bierhoff (2002), p. 118.
600 See Stich et al. (2010), pp. 200–201.
601 See ibid., pp. 172–174 and Tangney et al. (2007), p. 363.
602 Blair (2007), p. 387.

of what authorities say; the prototypical moral response pattern.[603] When questioned about the reasons for the wrongness of an act, they refer to the welfare of the victim less frequently than nonpsychopaths do. Rather, they tend to justify their 'moral' judgments by reference to social acceptability.[604] Blair believes that psychopathy might be due to a specific cognitive deficit that affects moral development. He posits the existence of a *violence inhibition mechanism* (VIM) similar to what presumably exists in some nonhuman animals such as dogs, which terminates aggressive behavior once an opponent has signaled submission. In his view, this VIM is a *prerequisite* for the development of the capacities for guilt, empathy, remorse, and sympathy as well as the ability to distinguish moral from conventional rules. Blair defines empathy as "an emotional reaction to a representation of the distressed internal state of another"[605], which is quite similar to Stueber's concept of affective empathy. Violations of moral rules involve infringements on the well-being of a victim, which have a particular significance only for those with a normally developed VIM. Without VIM, there is not that much of a difference in the emotional appeal of both moral and conventional rule transgressions, making the similarity in negative social responses to such transgressions relatively more salient. The VIM supposedly responds to observations of suffering with an impulse to withdraw from a violent confrontation. If that mechanism is deficient, distress cues will not inhibit violence. Even though this effect does not motivate aggression by itself, the resulting behavior will appear excessively cruel.[606] Interestingly, Blair's hypothesis does not locate the reason for the inability of psychopaths to distinguish moral from conventional norms at the level of empathic capacity in the sense of representing other people's mental states, but rather at a prior stage, namely the development of the VIM. He argues against the significance of the ability to represent other people's mental states (empathizing, role-taking) by reference to autistic children, who lack that capacity, but can nevertheless distinguish between moral and conventional rules.[607] Indeed, experiments indicate that psychopaths have a 'theory of mind'; at least they are not impaired in their ability to identify mental states by looking at other people's eyes.[608] Deficits in their moral behavior thus seem to result from a lack of interest in other people's well-being rather than from inability to

603 See Blair (1995), p. 20. Possibly, Blair's psychopathic subjects treated all of the rules as moral rules rather than treating all of them as conventional rules because they were incarcerated and eager to show improvement. See ibid., p. 23.
604 See ibid., pp. 18–20.
605 Ibid., p. 4.
606 See ibid., p. 11.
607 See ibid., p. 22.
608 'Theory of mind' refers to the ability to attribute mental states such as beliefs or emotions to others.

understand how they feel.[609] While Blair's VIM approach provides a good impression of how cognitive deficits might affect moral development, it fails to capture important aspects of empathy. For instance, the VIM is supposedly elicited in the context of violence and primarily motivates an end to the aggression. It is unclear how this process works in cases where an individual suffers for *other* reasons (e.g., because of an accident or if the suffering consists in sadness or similar psychological phenomena). Moreover, empathy and sympathy have been associated with motivations to help and actively improve the well-being of the target. Such behavior is not the focus of the VIM model.

Jesse Prinz, in contrast to many others, believes that empathy, understood as "a kind of vicarious emotion"[610] is neither necessary for moral judgment, nor for moral development, nor for motivating moral conduct.[611] Prinz presents David Hume as exponent of the position that empathy (sympathy, in Hume's terminology) is required for moral judgment. If it were not, pain and pleasure of others would not rouse us. In Prinz's account of Hume's position, empathy plays an epistemological role. Prinz considers several scenarios in which moral judgments occur, and concludes that none of them requires empathy: "As a descriptive claim it seems wrong to suppose that empathy is a precondition for moral judgment."[612] He argues that, for instance, deontological moral judgments often do *not* correspond with cumulative empathy; furthermore, the condemnation of transgressions without a salient victim, such as tax evasion, cannot easily be explained by reference to empathy. Prinz believes that other emotions or *sentiments* (his term for dispositions to feel emotions like anger, guilt, shame, etc.) as responses to specific types of actions can constitute moral judgments (the emotions do); no empathy is needed even if harm is involved. As for moral development, Prinz challenges Blair's hypothesis that empathy enables the development of a violence-inhibition mechanism and is therefore of crucial importance. In sum, he argues: "[P]sychopaths will lack emotions that facilitate moral education as well as the emotions that constitute moral judgments […]. Therefore, the deficit in moral competence can be explained without appeal to the empathy deficit."[613] Prinz points out studies that apparently, even though young children do experience empathetic responses, these responses do *not* figure in their moral considerations until they enter high school. In his view, the aforementioned emotions also provide motivation for moral conduct, again making empathy superfluous. In normative terms, Prinz claims that while empathy enables concern for those with

609 See Harris (2010), pp. 98–99.
610 Prinz (2011), p. 212.
611 See ibid., p. 213.
612 Ibid., p. 214.
613 Ibid., p. 218.

whom we have personal relations, it does not provide sufficient motivation and is too susceptible to bias to serve as a basis for *moral* concern. Making empathy the basis of a moral system would imply "preferential treatment and grotesque crimes of omission."[614]

So far, we have seen theories that take empathy to be essential for the development of morality as well as closely connected to the motivation of prosocial behavior, while Prinz argues that specific emotions, or dispositions to experience these emotions, can take over the explanatory roles occupied by empathy in these accounts. A related contrast exists regarding the role empathy plays in individual instances, rather than the development, of moral judgment.

What about the connection between empathy/sympathy and morality from the point of view of the instrumental/foundational framework developed in section 4.3.5? It seems to me that Prinz is partly correct in pointing out that at least in some cases, emotions that do not involve empathy determine moral judgment. In other cases however, empathy seems crucial. A look at the different moral concerns contained in MFT might help. Occurrences of anger elicited by violations of rules related to notions of fairness, harm, oppression, and the integrity of the in-group are good candidates for a foundational role of empathy. Why would we be vexed about the oppression, violation, or exploitation of others (in our group) if we were unable to grasp their situation (perspective taking) *and* have an emotional response to that perception that corresponds to the development of their well-being (affective empathy)? In this sense, empathic capacities are probably crucial to the establishment of some moral concerns, even if, once these foundations are in place, not every instance of moral anger need involve a strong empathic component (e.g., cases of principled anger). Considering the authority and sanctity foundations and the corresponding emotions, it seems plausible that they do not depend as much on empathic capacity because the entities which are to be protected (a hierarchical structure, a notion of purity) are not the kind of input for which the empathic psychological mechanisms evolved. Empathic mechanisms might be involved if, for instance, notions of sanctity depend on how some anthropomorphic goddess might feel about certain actions, or if the in-group is seen as a collective agent capable of emotional responses. Nevertheless, the connection seems weaker. Groups and goddesses do not provide facial, postural, or audible expressions of emotions the way our conspecifics or some nonhuman animals do. Without these inputs to empathic processes, the foundational relevance of empathy for the corresponding moral concerns is

614 Ibid., p. 227.

likely to be less significant. With regard to concerns for loyalty towards the in-group, empathy is presumably mediated by the delineation of this group. We are more empathic towards members of our in-group. Empathy, like yawn contagion, increases with familiarity and identification. Swiss soccer fans, for example, feel empathy only for supporters of their preferred club.[615] Researchers who try to understand violence often emphasize that aggression and lack of concern (for instance in times of war) is driven by dehumanization of the opponent.[616] Dehumanization excludes others from the in-group towards which empathic concern is appropriate. Biologist Sarah Brosnan reports that evidence for empathy in nonhuman animals is largely anecdotal, with the exception of studies of consolation behavior in nonhuman primates and responses to the distress of conspecifics in rats and mice.[617] In my view, however, caring for offspring the way mammals do already requires *some* empathic capacities. To the extent that concerns about harm, fairness in the sense of proportionality, and oppression are tied also to the emotional well-being of individuals, empathic capacities promote adherence to a moral code by enabling us to perceive and anticipate the effects of actions on others and direct our behavior and the behavior of others accordingly. Moreover, they can provide the (intrinsic) motivational impulse to do so.

According to many accounts, being able to imagine what one would feel like in the place of others is crucial to individual moral development. From feeling with others, we learn whether specific behaviors are morally permissible. For instance, understanding the situation another is in and simulating one's own emotional state in that situation is one way to be moved by another's predicament. Another is to respond directly to the other person's emotion, either supported by conscious processing, or based on more automatic mechanisms, possibly involving mirror neurons.[618]

## 4.5 Self-Conscious Moral Emotions

### *4.5.1 Self-Critical Emotions*

*Self-critical* moral emotions such as guilt, shame, and embarrassment generate motivation to behave in accordance with established social norms and thereby prevent individuals from becoming objects of other-critical emotions, or weaken such emotions if already present.[619] Humans have a strong desire to belong to groups and a corresponding tendency to comply

615 See De Waal (2013), p. 138.
616 See for instance Bandura (1999).
617 See Brosnan (2014), pp. 94–96
618 See Kitcher (2011), p. 25.
619 See Haidt (2003c), p. 859.

with default behaviors in those groups. A general motive to imitate the behavior of group members serves as a heuristic for appropriate conduct.[620] Failures to conform elicit different emotions contingent on their subject matter/domain, seriousness, and permanence not only in others, but also in the transgressor herself. Evidence for psychological phenomena similar to shame and guilt in related species provide reasons to think that experiences of shame and guilt do not necessarily require the presence of more or less formally established, explicit social norms. To the extent that this is the case, these emotions potentially have a foundational effect on the moral domain by marking specific eliciting events as relevant. Whether such an effect is plausible also for embarrassment is less clear.

#### *4.5.1.1 Shame and Embarrassment*

The evolutionary roots of shame and embarrassment (and their positively valenced counterpart, pride) presumably lie in mental abilities that regulate dominant and submissive behavior in hierarchies. So far, it is not certain whether shame or embarrassment is phylogenetically prior. My impression is that shame is more frequently discussed, maybe because it is a more powerful emotion. Nevertheless, the evolutionary rationales for its development also apply to embarrassment to the extent that both emotions are similar in terms of elicitors and action tendencies. Full-blown shame presumably has its origins in a capability to experience 'protoshame', which motivates status-adequate conduct towards higher-ranking individuals, for example unobtrusiveness or lowering one's gaze.[621] Since allocation of scarce resources like food, mates, and territory is an important function of hierarchy, protoshame makes individuals avoid quarrels over such resources in acceptance of the other's superiority or motivates appeasing, submissive behavior. Similar patterns are present in humans: Shame and embarrassment attenuate motivations to raise claims of any sort; moreover, we are less likely to experience them in the presence of lower-ranking people. Unlike humans, however, hierarchical animals like wolves seem to experience similar sensations only when caught in the act by a superior.[622] Tracing the origins of these emotions back to processes that regulate behavior in potentially dangerous dominance negotiations explains why shame often goes along with motivations to flee or hide.[623] This evolutionary account gains plausibility in light of the finding that shame apparently encompasses experiences of fear and respect to a larger degree in some non-WEIRD (Western, Educated, Industrialized, Rich, and Democratic)

620 See Gigerenzer (2007), p. 220.
621 See Haidt (2003c), p. 860.
622 See Boehm (2012), p. 23.
623 See Fessler (2010), p. 92.

societies, while Western cultures often see shame as primarily related to personal and moral worth of the self. Fear and respect have important functions in regulating subordinance relations.[624] They are also important for the establishment of punishment systems that enable advanced forms of cooperation. The perception of hierarchy prevents individuals from intervening if a higher-ranking member of their group punishes an ally, making systems of punishment less costly and more reliable.[625] Interestingly, words expressing shame-like concepts exist in almost every culture, while many languages, including those of foraging tribes, lack a word for 'guilt'.[626]

Embarrassment occurs when we become aware of having violated a social norm and temporarily reduces one's aplomb in social interactions, particularly if others appear to notice the transgression.[627] It has been defined as "an aversive state of mortification, abashment, and chagrin that follows public social predicaments"[628] and is marked by a rather unique physiological response: increased blood flow in subcutaneous capillaries of the face (blushing).[629] At least in conjunction with this specific physiological mark, it does not occur in nonhuman apes: they do not blush.[630] In human children, both blushing and feelings of shame emerge at around two years of age.[631] Embarrassment generates tendencies to behave in agreeable and conciliatory ways; such behavior may also result from a motivation to avoid embarrassment caused by nonconformity (anticipatory effect).[632] Shame, in contrast, corresponds to deficits of the self in various domains (morality, aesthetics, or competence) which the person feeling ashamed perceives as rather grave and permanent. Shame (like guilt) is related to transgressions of moral norms, while embarrassment tends to arise in response to violations of other social conventions.[633]

As the linkage with moral rather than conventional transgressions suggests, shame is more intense than embarrassment. Shame involves a motivation to escape from the situation that triggered it and is associated with deferent posture and self-concealment.[634] There are slight differences in focus between the ways in which researchers distinguish the elicitors of shame and guilt: Some argue that, while guilt typically arises in response to a specific,

624 See ibid.
625 See Kitcher (2011), p. 89.
626 See Boehm (2012), p. 20.
627 See Moll et al. (2008b), p. 14.
628 Tangney et al. (2007), p. 359.
629 See Schacter et al. (2009), p. 373.
630 See Prinz (2012), p. 321, Boehm (2012), p. 120.
631 See ibid., p. 223.
632 See Tangney et al. (2007), p. 360 and Haidt (2003c), p. 840.
633 See Tangney et al. (2007), p. 359.
634 These behaviors can be triggered by increases in the levels of certain proteins and hormones. See ibid., p. 350.

negatively evaluated *action* of the self, shame contains an appraisal of the *self* as seriously and permanently deficient.[635] Referring to Shweder's three ethics, other authors associate shame with crimes against nature (divinity) and guilt with crimes against persons (autonomy): On that account, the specific objectionableness of the behavior determines which emotion is elicited.[636] In the context of such an account, Jesse Prinz suggests that embarrassment is phylogenetically older than shame, and that shame arose only once humans developed notions of a natural order.[637] For my purposes, it is not crucial whether and to what extent protoshame and embarrassment are identical. The differences in emphasis researchers put on different facets of the elicitors (domain or action vs. character flaw) might well be compatible; they also point to the leeway individuals have in interpreting their own actions. Different individuals may construe similar actions in various ways. A murderer or violent felon could be ashamed of his crime if he construes it as a transgression against the sacredness of human life that belies fundamental flaws in his character, or he might (as well) experience guilt if he interprets his act as having been 'out of character'. Shame is associated with real or imagined audiences to the deficiency.[638] Generally, it seems that shame can have severe negative consequences for the person experiencing it (it has also been found to worsen immune function)[639], while guilt often leads to adaptive behavior (for instance, through inhibition).[640] Displays of shame appear to decrease the degree to which groups that suffered from serious injustices feel insulted by offers of compensation.[641]

What role do shame and embarrassment play in morality? Instrumental effects are not far to seek: Shame motivates to cease the offensive activity or at least avoid the attention of others. The effects of embarrassment are similar, but less intense and of shorter duration. The anticipation of both shame and embarrassment may lead us to learn about the rules and customs of the people we deal with, and keep us from engaging in inappropriate behavior in the first place. With respect to shame, we can at least speculate that it has some foundational influence on morality as well. To the extent that shame (and, to a lesser degree, embarrassment) stems from psychological mechanisms that enabled our ancestors to navigate group hierarchies, attentiveness to social roles and the obligations and privileges they bring might be part of our evolutionary heritage, and a central feature of many moralities.

635 See ibid., p. 349.
636 See Prinz (2012), p. 307.
637 See Prinz (2007b), p. 78.
638 See Tangney et al. (2007), p. 349.
639 See ibid., pp. 356–357.
640 See ibid., p. 350, and Giner-Sorolla et al. (2010), p. 91.
641 See ibid.

#### *4.5.1.2 Guilt*

Guilt is often considered a prototypical moral emotion. It arises primarily when we cause harm to members of our in-group (others about whose well-being we care), its intensity corresponds to depth of attachment and severity of harm.[642] Unlike shame and embarrassment, guilt is associated not with motivations to withdraw, but rather with a willingness to 'make it up' to the victim or her group by apologizing, confessing, paying reparations, self-criticizing, or whatever other practice is deemed appropriate.[643] It also generates a readiness to accept sanctions that otherwise might be perceived as an unjustified aggression and met with anger and retaliation. In human children, guilt emerges later than shame, between the ages of five and nine.[644] Individuals differ in guilt-proneness.[645] As do other emotions associated with prosocial action tendencies, the experience of guilt appears to involve the ventromedial prefrontal cortex.[646] While shame is elicited by manifestations of more general personality deficits, guilt relates to specific harmful events.

Jesse Prinz discusses whether guilt is a 'disembodied' emotion, i.e., an emotion without a bodily component.[647] Unlike me, he considers guilt phylogenetically recent. At first sight, it might appear that guilt is *not* marked by typical physiological responses. However, Prinz cites an informal study of his in which frowning was recognized as the facial expression most appropriate to a feeling of guilt. Moreover, he argues that the fact that we would be suspicious of someone who claims to feel guilty but does not show any change in the way he carries himself shows that we have some notion about how guilt is typically expressed, for instance by avoiding the eyes of those you have harmed or betrayed. Even if it may be hard to pin down a specific physiological response typical of guilt, there are some indications that a psychological phenomenon similar to guilt is present in nonhuman primates, and possibly other mammals.[648] The claims to similarity are based on elicitors and action tendencies: Frans de Waal reports that chimpanzees regularly engage in 'reconciliation' after conflict: "10 minutes after a fight, one male may hold out a hand to the other invitingly, leading to embracing and kissing, followed by mutual grooming."[649] This is quite different from behavior produced by the avoidance motivation associated with shame. Those who

642 See Prinz & Nichols (2010), p. 133.
643 See ibid., p. 134.
644 See TenHouten (2009), p. 51.
645 See ibid., p. 96.
646 See Moll et al. (2008a), p. 168.
647 See Prinz (2007b), pp. 59–60.
648 But see Boehm (2012), pp. 124–125.
649 De Waal (2004), p. 19. De Waal agrees that these behaviors can be interpreted as resulting from psychological predecessors of guilt (personal communication, Forum Scientiarum Tübingen, June 2012).

feel guilty take constructive measures: They confess, apologize, and try to undo the damage they caused. Those who are ashamed get defensive and attempt to increase the distance between themselves and others.[650] Since guilt motivates actions that aim to improve the other person's well-being, it is associated with increased empathy in the sense of trying to find out how one can make the victim feel better again.[651] Shame, on the other hand, appears to inhibit the ability of individuals to put themselves in other people's shoes.[652]

Guilt seems hardly to appear in nonmoral contexts. If it does occur in nonmoral contexts, it often seems like a kind of mistake in which someone inappropriately takes a moral stance (e.g., feeling guilty about breaking a diet). Guilt is closely associated with the idea of 'conscience'. Sometimes, those who are better off compared to others experience guilt even if it is unclear whether the inequality was caused by a transgression. 'Survivor guilt', as sometimes experienced by individuals who live through a traumatic event while others perish, is a prominent example of this phenomenon. Possibly, it occurs because subjects mistakenly believe that they had some degree of control over harm suffered by others, which generates a sense of responsibility. Actions for which the agent is responsible and that hurt someone 'near and dear' are prototypical elicitors of guilt; omissions are less pertinent.[653]

Prinz proposes the following core relational theme for guilt: "[S]omeone I am concerned about has been harmed and I have responsibility for that in virtue of what I have done or failed to do."[654] Occurrences of shame, in contrast, appear not to require the same sense of control and responsibility associated with guilt; people might even feel ashamed of compulsive behavior.[655] The distinction between shame and guilt mirrors some aspects of the differences between anger and disgust. One can feel guilty for an act without believing that one is generally a person of bad character, just as one can be angry with somebody else in response to some specific act without believing that that person is morally deficient. Shame and disgust, in contrast, are felt with respect to agents rather than actions. Another way to put this is to say that actions that elicit shame or disgust cause that emotion to be associated with the agent.[656] Remember that according to the CAD-triad hypothesis, anger is associated

650 See Tangney et al. (2007), p. 350.
651 See Bierhoff (2002), p. 140: A questionnaire study with undergraduates found correlation between empathy (perspective taking, empathic concern, fantasy) and guilt.
652 See Tangney et al. (2007), p. 350.
653 See Prinz & Nichols (2010), p. 134, De Hooge et al. (2011), p. 464.
654 Prinz & Nichols (2010), p. 134. Remember that core relational themes are emotion-eliciting organism-environment relations.
655 See ibid., p. 135.
656 See ibid., p. 136.

with violations of the ethics of autonomy (Shweder's framework), which translates to transgressions in the domains of oppression, harm, and fairness in Haidt et al.'s MFT. One might say that guilt is a self-directed counterpart of anger, and similarly related to these domains. Could this thought be extended to associate shame with violations of divinity/purity norms? Such a patterning seems at least intuitively plausible. In a similar vein, Jesse Prinz argues that anger and guilt are elicited by crimes against persons, while disgust and shame are reactions to crimes against nature (natural/sacred order).[657] Interestingly, while the CAD-triad hypothesis maintains that transgressions against the ethics of community (violations of rank/hierarchy or in-group integrity) performed by others are typical elicitors of contempt, it is not quite clear what the emotional counterpart is on the transgressor's side, and whether it is distinct from shame and guilt. Prinz merely mentions "some kind of self-loathing"[658]. Both shame and guilt can be experienced vicariously, elicited by actions performed by members of one's group. Paralleling the nonvicarious variants, group-based shame arises when positive group identity is threatened, while group-based guilt arises when both harm done to another group or individual *and* a relation with the responsible agent are salient.[659]

Early in the twentieth century, Freud understood guilt as resulting from conflicts between the id and the superego, or between the interests of the individual and the requirements of existence in a society. During that period, negative *intra*personal effects of guilt (e.g., sadness and neuroses) were the focus of attention.[660] More recently, researchers emphasize the *inter*personal, positive functions of guilt:[661] In social dilemma games, for instance, participants who feel guilt behave more prosocially.[662] Individuals who are prone to feelings of guilt are less likely to engage in risky or antisocial behavior.[663]

In a constellation similar to Milgram's obedience experiments (see section 1.3.3), 'teachers' were asked to make phone calls on behalf of an environmental organization. When asked by someone who *witnessed* their administration of electric shocks, subjects were more likely to make the calls than teachers asked by the learner or others who were asked by the learner, but did not administer electric shocks. Apparently, the subjects were not motivated to make amends to the learner, but make some impression on the witness. Moreover, guilt seems most likely to induce prosocial behavior when the transgressor has had no other

657 See Prinz (2011), p. 215.
658 Ibid.
659 See Tangney et al. (2007), p. 359.
660 See Prinz & Nichols (2010), p. 132.
661 See De Hooge et al. (2011), p. 462.
662 See ibid., p. 463.
663 See Tangney et al. (2007), p. 354.

outlet for his guilt.[664] Apart from its desirable aspects, guilt can also have less attractive effects beyond the aversive character of its experience: In some cases, guilt-induced preoccupation with finding a way to repair the damage done can make individuals neglect other social relationships.[665] These results point to a deficit of current research on the behavioral tendencies of specific emotions: In most cases, these emotions are investigated in paradigms that involve one or two individuals. A comprehensive picture of the functions and side effects of emotions requires research that extends to three-person interactions.[666]

Ontogenetically, the propensity to feel guilt might stem from withdrawal of love by parents in response to undesirable behavior, which causes sadness (remember that the core relational theme of sadness is the loss of something precious). While sadness is first elicited by the loss of affectionate behavior of others, it becomes associated with the transgressions themselves.[667] This developmental hypothesis explains why guilt is closely related to harm, because harmful transgressions are likely to damage relationships. There might be differences between cultures with respect to the prominence of the concept of guilt: It is very salient in North America and generally with Christians and Jews, less so with Buddhists, Confucians, Hindus, and Muslims.[668] The development of the capacity to experience guilt is related to the emergence of affective empathy. Both require the ability to realize that somebody else is in distress and motivate helping behavior. Only guilt, however, includes the appraisal that the self is responsible for the victim's state.[669]

Like the aforementioned emotions, guilt has a twofold motivational function: Individuals make amends (confess, seek punishment, etc.) once they feel guilty, but they may also anticipate the unpleasant experience of guilt and thus not perform actions that they believe they would feel guilty about in the first place. These anticipatory effects might extend beyond the effects of anticipated anger: Guilt does not always require detection, thus it could prevent transgressions even if the transgressor would have gotten away with it. This is also true for actualized guilt: Guilt-ridden transgressors might confess and make amends for transgressions even if they had not previously been identified as culprits.[670] It is conceivable that guilt is a more powerful contributor to conformity with norms than anger, because the anticipation of anger works through fear, which is a relatively weak motivator for moral

664 See Prinz & Nichols (2010), pp. 138–139. For instance, people donate more before making a confession than afterwards.
665 See De Hooge et al. (2011), p. 471.
666 See ibid.
667 See Prinz & Nichols (2010), p. 136, Prinz (2007b), p. 78.
668 See Boehm (2012), p. 19.
669 See Bierhoff (2002), p. 139.
670 See Prinz & Nichols (2010), p. 141.

behavior. Ensuring compliance trough fear is less effective in moral education than love withdrawal, which generates guilt.[671] In this context, it is interesting that there seems to be a (small) gender gap in guilt-proneness, with women feeling guilty more often and more intensely than men do.[672] At the same time, females account only for a small percentage of the prison population in most countries. One explanation for this phenomenon is that guilt prevents bad behavior, in addition to promoting prosocial behavior in the sense of making amends to victims and helpfulness towards others.

### *4.5.2 A Self-Praising Emotion: Pride*

Pride is the counterpart of shame in terms of how it regulates behavior in hierarchies and groups. It increases dominant behavior and is associated with greater motivation to persevere.[673] These observations lend themselves to an evolutionary explanation: Pride might derive from the psychological systems that motivate striving for power and status in animals that live in hierarchically organized groups. By themselves, these motivations have nothing particularly moral about them. However, pride can be affected by what one's social surroundings value. We take pride in achieving things that our contemporaries appreciate. If they honor norm conformity or moral excellence, pride or the desire to be proud of oneself can motivate compliance with these social preferences and thus contribute to prosociality, norm stabilization, and possibly overall well-being.[674]

Because pride can take expectations of others into account, it might be understood as the positive counterpart of guilt or shame. Tagney has suggested a distinction between self-related *alpha*- and behavior-related *beta*-pride, according to which alpha-pride is the positive pendant of shame, while beta-pride corresponds to guilt. Ekman proposes the term *fiero* for the positive emotion we feel when we accomplish a challenging task, and whose experience does not require an audience to this achievement; the concept seems very similar to beta-pride.[675] Others have named similar concepts hubris and (achievement-oriented) pride, respectively and provided empirical evidence that what is loosely referred to as pride are in fact two distinct phenomena.[676] Just like guilt, (beta-)pride seems to require a certain sense of control over and responsibility for being an esteemed person or the producer of a socially valued outcome.[677] If one saves a life unintentionally, pride seems inappropriate.

---

671 See ibid., p. 137.
672 See, for instance, Else-Quest et al. (2012), p. 964.
673 See Valdesolo & DeSteno (2011), p. 277.
674 See ibid.
675 See Ekman (2003), p. 196.
676 See Tangney et al. (2007), p. 360.
677 See ibid.

If there are foundational effects of beta-pride, they probably consist in establishing a concern for status and hierarchies similar to shame, but this time, as it were, with a view from the above, rather than from below. There are instrumental effects, both based on an appreciation of what is considered morally commendable behavior, and from positive social effects of individual ambition in which concerns for the welfare of others do not play an important role. The latter effects certainly depend on the presence of institutions that confine the roads to success in a way that benefits society. Presumably, emotions that *establish* moral concerns (foundational effect) regularly *promote* behavior corresponding to that concern at least to some extent.

# 5 Models of Moral Cognition: The Interplay of Intuition, Emotion, and Reason

The previous chapter explained how important emotions are in establishing fundamental moral concerns, but also in bringing about morally commendable states of the world. However, these explanations do not quite seem to address the skepticism regarding some moral judgments spurred by the experiments of Greene and Haidt. Singer pushes another concern to the fore: Allegedly, some judgments are insufficiently determined by 'reasoning'. Greene makes similar, but more detailed recommendations to arrive at reliable moral evaluations: At least in fundamentally new situations, moral judgments should be generated using system 2. In order to evaluate the extent to which these proposals can and should be implemented, it is important to consider the nature of the psychological mechanisms involved in moral judgment in detail. I have already emphasized the seminal influence of emotion-elicitor relations on notions of moral relevance. Such moral-emotional responses to elicitors are often intuitive: They proceed automatically, quickly, and effortlessly; only their outputs, but not the processes involved in generating them, enter consciousness.[678] This 'intuitiveness' is probably inherited from evolutionary predecessors of modern humans which had neither time nor capacities to ponder their reactions for long. However, not all moral judgments are equally intuitive. Some require tedious reasoning (particularly among philosophers), for example, when emotional intuitions are ambiguous. Whether and how emotions and reason interact in moral judgment is a perennial debate in philosophy, and has now become a focus of research into the mental processes underlying moral functioning. These efforts have yielded several models of moral judgment, some of which I discuss in this chapter to create a clearer picture of the morally efficacious and its relation to the morally relevant. Moreover, understanding the cognitive processes corresponding to broad labels like 'reason' and 'emotion' promises a more precise discussion of demands for 'more reasoned' moral judgment. The mechanics of moral judgment are important not only to the evaluation of Greene's position, but also of the other views on the normative and metaethical significance of moral psychology. While the emotional character of some moral judgments is one concern, reliance on intuitive mechanisms and evolved psychological processes have also called problematic. These topics are related: Emotions are, at least to some extent, ancient evolved psychological mechanisms; emotion processing is often intuitive.

[678] See Haidt (2001), p. 818.

Many researchers think of moral intuitions as heuristic processes that do not grasp 'the real thing'. Such normative conceptions of how proper moral judgment should come about depend crucially on what the human mind is actually capable of, and on whether the mechanisms whose reliability they doubt and those whose use they propose are actually distinct.

This chapter further explores the mechanics of moral judgment. In particular, it discusses the roles of emotion, intuition, and reason, and investigates how fundamental concerns enter the judgment process. There are many models of how supposedly distinct mental processes interact in moral judgment. Like the dual-system model of cognition, many of them distinguish rational from emotional or intuitive capacities. They differ, however, in the roles or relative importance they ascribe to these systems. I will not provide a comprehensive overview of judgment models, but rather highlight points of contention between the well-known models of Greene and Haidt, introduce the linguistic analogy in moral psychology as third important suggestion, and argue that there are many ways to arrive at moral evaluations.

Many authors assume that the human mind is capable of two alternative, but interacting modes of processing information often referred to as system 1 ('experiential mode') and system 2 ('rational mode'). These dual-process models of cognition are commonplace in contemporary psychology. System 1- and system 2 processes differ on several dimensions: System 1 processing is automatic, largely unconscious, fast, effortless, often associated with emotions, and possibly executed by multiple mental modules.[679] System 2 processing is relatively slow, controlled, conscious, effortful, and modular to a lesser extent. System 2 is more recent and associated with the phylogenetically youngest brain structures, mainly the neocortex, while functions performed by system 1 are presumably located in phylogenetically older structures like the limbic system. Both systems are frequently active concurrently, system 2 is likely to be anchored on system 1 processes, and their outputs sometimes compete.

How do we find out whether a given task is processed by system 1 or system 2? Supposedly, overall capacity for system 2 processing is limited. Effortful processes therefore disrupt each other, while intuitive processes do not. If the processing of a task is slowed down under 'cognitive loads' such as the requirement to memorize random numbers, that task is presumably at least in part processed by system 2. The relative extent of system 1 and 2 processing involved in a given task depends both on individual and task characteristics. System 1 is arguably the default processing mode and controls behavior in many situations,

[679] See Liu & Hao (2011), p. 204.

without us being aware of it. It is helpful to compare intuitive system 1 processing and deliberate system 2 reasoning to perception: Like reasoning, intuitions can occur in response to concepts stored in memory, while perception processes only current input. In terms of the immediacy of its output, however, system 1 is very similar to perception, while system 2 is not.[680]

## 5.1 Haidt's Social-Intuitionist Model of Moral Judgment

Since chapters 3 and 4 lean quite heavily on the related Moral Foundations Theory and because of its prevalence in moral psychology, I consider Jonathan Haidt's *social-intuitionist* model of moral judgment (SIM) first. The name indicates its two most distinctive features: Judgment is shaped by intuition (system 1), not reasoning, and it is frequently subject to social influence.[681]

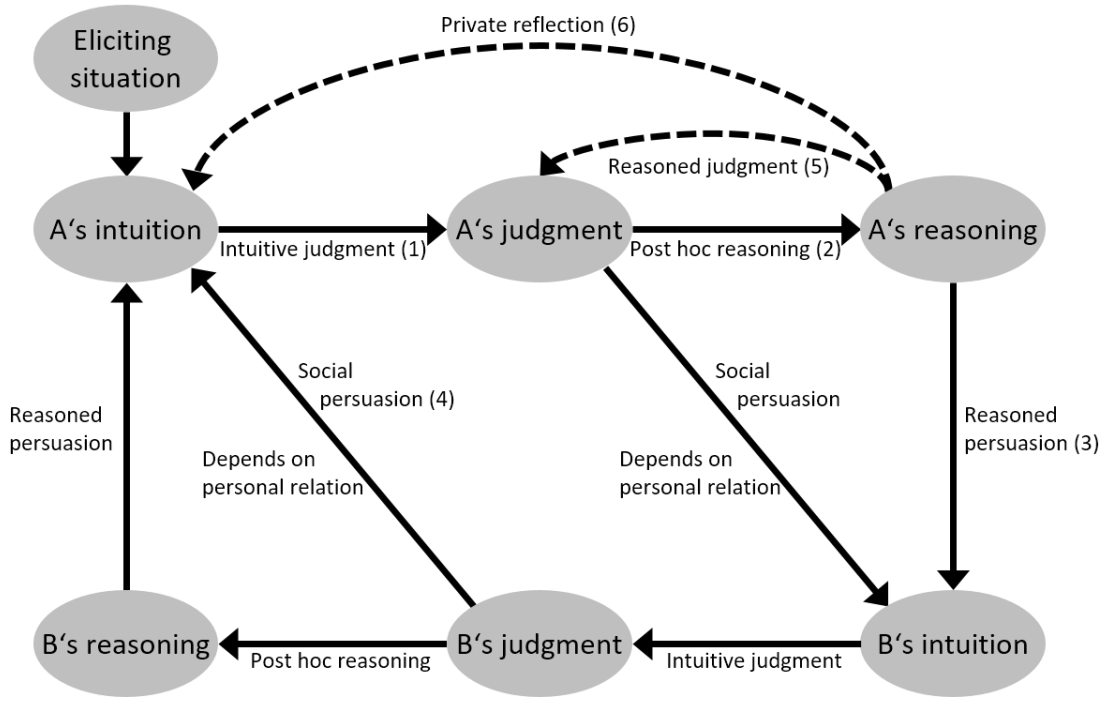


**Figure 4: The Social-Intuitionist Model of Moral Judgment**

From Haidt, J. (2001): The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. In: *Psychological Review*, 108 (4), pp. 814–834, p. 815,
published by the American Psychological Association, adapted with permission

680 See Kahneman & Sunstein (2005), p. 93. There is some evidence that not all system 2 processing involves conscious control. For instance, we seem to be able to suppress implicit attitudes that conflict with reflectively endorsed attitudes without being aware of that suppression. See Kennett & Fine (2009), p. 92.

681 See Haidt (2001).

Since intuitions have an affective quality, the intuitive response to an eliciting situation and link 1 connect this model with the fundamental influence of emotions captured in MFT.[682] Conscious, verbal reasoning (link 2) resembles a lawyer or press secretary, justifying judgments after the fact.[683] As mentioned in section 1.3.2, Haidt defines moral intuition as

> […] the sudden appearance in consciousness, or at the fringe of consciousness, of an evaluative feeling (like–dislike, good–bad) about the character or actions of a person, without any conscious awareness of having gone through steps of search, weighing evidence, or inferring a conclusion.[684]

Since the process of intuitive judgment formation is unconscious, reasoning draws on "culturally supplied norms for evaluating and criticizing the behavior of others. […] (e.g., 'unprovoked harm is bad'; 'people should strive to live up to God's commandments')"[685]. While these justifications *may* refer to the judgment's origin, this is not necessarily the case, and might even be unlikely.

Reasoning can lead to judgment revision, either by "sheer force of logic"[686] (link 5) or by altering the perception of the situation and thereby triggering new intuitions (link 6). While Haidt acknowledges these possibilities, he believes that most justifications rationalize the initial intuition.[687] Links 3 and 4 represent social determinants of moral judgment. We may align our intuitions with the verdict of esteemed or likeable others (link 4). In other cases, not the judgment itself, but the *justifications* others give affect our intuitions (link 3). Pondering a complex moral conundrum on one's own can be modeled as repeated engagement of links 6, 1 and 2: For instance, we can make a conscious effort to take the perspective of all affected parties in turn and thereby elicit various intuitions. If one intuition gains

682 Haidt discusses an empiricist understanding of the origin of intuitions. On this view, intuitions develop according to reinforcement (behaviorism) or learning from parents, peers, and media (social learning theory). While such processes occur and are important, Haidt argues against the view that moral intuitions depend *exclusively* on external influences. Children regularly resist the values they are taught, they also do not learn all norms with similar ease (no equipotentiality). See Haidt & Bjorklund (2008), p. 201.

683 See Haidt (2001), p. 818, Nado et al. (2009), p. 627.

684 Haidt & Bjorklund (2008), p. 188. The case of medical doctor Bernard Nathanson provides a particularly vivid example of how sensory impressions, rather than argument, radically alter moral views. Nathanson was involved in the founding of the National Association for the Repeal of Abortion Laws (a prominent pro-choice organization) and himself a provider of abortions, but became a prominent pro-life activist after observing abortions through ultrasound. As an obstetrician, he knew very well what an abortion involves. His opinion only changed when he could *see* the process (including what he perceived as indications of pain in the fetus). See http://en.wikipedia.org/wiki/Bernard_Nathanson.

685 Haidt (2001), p. 822.

686 Ibid., p. 819.

687 See ibid., Nado et al. (2009), p. 627. Frequencies may differ for people with special training, e.g., moral philosophers.

dominance, or if several intuitions can be reconciled with each other, the process terminates. The corresponding judgment then 'feels' acceptable or justified.[688] The social-intuitionist model can explain decouplings of behavior and judgment: Moral argument serves to justify action and convince others (even if the action was originally self-serving).[689]

## 5.2 Greene's Dual-Process Model of Moral Judgment

Social-interactive influences were less prominent in Joshua Greene's earlier publications. Rather, his dual-process model focused on the link between certain features of the iudicandum and the psychological mechanisms that form the judgment. Personalness of harm (possibly in the sense of personal force), for instance, triggers alarm-like emotions that generate the impression of incompatibility of the act with nonoffsettable moral concerns that Greene considers typical of deontology. The processing of impersonal harms, in contrast, involves rational capacities and/or currency-like emotions. These capacities *weigh* different concerns; a cognitive style associated with consequentialist/utilitarian morality.[690] With respect to impersonal cases, Greene grants reason a more prominent role than the SIM does.[691]

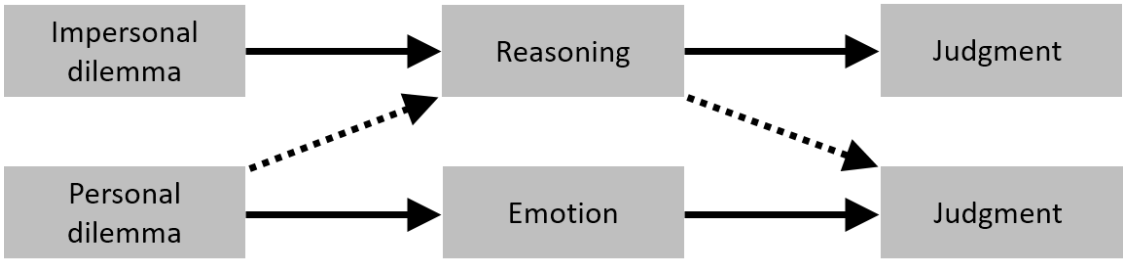


**Figure 5: Greene's Model of Moral Judgment**

From Nado, J., Kelly, D. R. & Stich, S. (2006): *Moral Judgment (final draft)*. http://www.jennifernado.net/pdfs/papers/NadoKellyStichMoralJudgment2007.pdf (accessed 17.06.2017)

According to Greene's model, this is what happens in the processing of trolley problems: In the footbridge dilemma, personal force or some other aspect of 'personalness' activates evolved, alarm-like emotional mechanisms. In a few individuals, reasoning overrides the emotional response (dotted arrows in Figure 5); these individuals consider pushing the stranger permissible. 'Impersonal' dilemmas like 'switch' do *not* cause alarm-like emotional responses; thus, conscious reasoning, possibly based on currency-like emotions, shapes the

688 See Haidt (2001), p. 829.
689 See Haidt & Bjorklund (2008), p. 211.
690 See Paxton & Greene (2010), p. 513.
691 See Nado et al. (2009), pp. 628–629.

corresponding judgments. In recent work, Greene incorporates social influence and points to two important remaining differences between the social-intuitionist model and his dual-process model. Firstly, he believes that reasoned, particularly consequentialist, judgment arrived at independently (i.e., not based on social or reasoned persuasion) is more frequent than Haidt assumes. Secondly, he claims that the reasoning of others can affect moral judgment *without* triggering new intuitions.[692] Rather than argue about the relative frequency of reasoned and emotion/intuition-based judgment, Greene wants to show that "genuinely reasoned moral discussion" is possible.[693] Such discussion can socially generate *counterintuitive* judgment, an option not captured in Haidt's model. This is a reasonable suggestion. According to the SIM, we are more likely to change our views as an effect of the reasoning of friendly *others* rather than our own deliberation, since reasoning supposedly developed to defend *our own* intuitions, and influence others accordingly.[694] This, however, does not explain why counterintuitive reasoned judgment should be possible privately (link 5 in Figure 4), but not socially. In the SIM, social influence *always* takes effect by changing intuitions. There is also a disagreement between Greene and Haidt about whether 'reasoned persuasion' (link 3), in which other people's arguments influence an individual's intuitions, should count as reasoning. While Haidt categorizes such persuasion as reasoning process and accordingly claims that reasoning does play an important role in his model, Greene argues that, in the language of his camera analogy, the other's arguments in fact engage the targeted individual's automatic settings rather than the manual mode, which is why 'reasoned persuasion' should *not* count as reasoning.[695]

More precisely, Greene is concerned not with factual inferences, but reasoning understood as assessing the *consistency* of one's moral commitments, be they general principles or particular judgments.[696] He presents a collection of evidence for the claim that not all (socially induced) change in moral attitudes is based on changes in intuitions, but that people do at least sometimes consciously apply moral principles and override the prepotent intuitive response. Haidt had previously defended his model against the claim that stronger activation of 'rational' brain areas (DLPFC) in difficult moral judgments such as 'crying baby' suggest that these decisions are not affect-driven: "[T]here is indeed a conflict between potential responses, and additional areas of the brain become active to help resolve this

692 See Paxton & Greene (2010), p. 514 and Greene (2014b), p. 335.
693 See Paxton & Greene (2010), pp. 514–515.
694 See Haidt (2012), p. 68, Baumeister (2005), p. 240.
695 See Greene (2014b), p. 385.
696 See Paxton & Greene (2010), p. 516.

conflict, but ultimately the person decides based on a feeling of rightness, rather than a deduction of some kind."[697] In order to address this defense, Greene quotes research on so-called 'implicit attitudes' that occur without conscious awareness: A majority of white people has negative implicit attitudes toward black people, while their *explicit* attitudes are neutral or positive. Greene argues that these explicit attitudes do *not* correspond to intuitions, but rather result from a conscious effort to do the right thing *in spite of* possible unconscious intuitions (implicit attitudes) which point the other way. If this conjecture is an adequate interpretation of the processes behind some episodes of judgment change, explicit attitudes are sincere, and the change is triggered by social influence, it might indeed describe a mechanism *not* captured by the social-intuitionist model, since it is not based on a change in intuitions.[698] On the other hand, it is also conceivable that one intuition (the original implicit attitude) is overridden by a newly formed, stronger intuition. I address possible transition of attitudes from system 2 to system 1 in more detail in chapter 5.5.

Philosophers Ron Mallon and Shaun Nichols argue that intuition-emphasizing dual-process models of moral cognition neglect the possibility that we sometimes apply *consciously* available moral rules *effortlessly*. In their view, this option speaks against the assumption that emotional, intuitive, *unconscious* system 1 processes dominate moral judgment. Since no one has so far assessed the relative frequency of system-1- and system-2-dominated judgments in the real world (the 'counting problem'), proponents of system-1 dominance often refer to the claim that system 2 processing is effortful, and that cognitive resources are limited. Therefore, they argue, the majority of moral judgments have to result from system 1 processing. If, however, some *consciously* available moral rules can be applied *without* effort, this argument fails, and it remains unclear whether emotional-intuitive judgment is in fact more frequent.[699] Is this a decisive argument against intuitionist positions?

Proponents of system 1 dominance do not have to claim that emotions and intuitions are completely unconscious. Rather, the *output* of these processes is often conscious, but the *preceding* processing is not. From this point of view, the kind of effortless rule-application Mallon and Nichols describe could be a subtype of system 1 processing, possibly of low emotional intensity. Would this be a problematic concession to make for those who hold that emotional intuitions have a fundamental role in shaping morality? I do not think so. I am not arguing that *strong* emotional activation is involved in *every* moral judgment. The idea

[697] Haidt & Bjorklund (2008), p. 195.
[698] See Paxton & Greene (2010), pp. 523–524.
[699] See Mallon & Nichols (2011), p. 285.

is, rather, that such consciously applied rules have to result from considerations that resonate with emotionally anchored fundamental concerns; otherwise, they would not stabilize. The point made by Nichols and Mallon is nevertheless useful, because it highlights the importance of specifying exactly what one is talking about: The execution of everyday moral judgment, or the *establishment* of the rules and mental habits that guide these ordinary, low-conflict judgments.

Note also that the dialectic constellation is similar to the debate surrounding the moral/conventional distinction: Researchers assume a nomological clustering of properties and assign labels to these clusters. Then someone takes a closer look and finds that the clustering is less strict than supposed; that the individual properties can vary in more permutations than suspected. While in the moral/conventional debate it was assumed that *moral* judgments are *always* serious, universally valid, authority independent, and elicited by and justified with respect to harm, rights and justice, the distinction of system 1 and system 2 at least suggests (if it does not hold explicitly) that *all* mental processes which are conscious are also effortful, slow, etc. Reality might be more complex.

Which understanding of moral judgment should we extract from these models? It seems that disagreements between them do not concern the question whether system 1 and system 2 are *at all* efficacious in moral judgment. Differences in emphasis between the models of Greene and Haidt might result from a *focus on different phenomena*: While Haidt aims to depict quotidian judgment; Greene often seems more interested in the weight different processes have in the accomplishment of moral change or progress.

I believe it makes sense not to think of morality and moral judgment as unified phenomena. Given the rich evolutionary and cultural history of what common parlance can take to be part of morality, attempts at reduction will most likely engender fruitless debates about the 'correct' model of moral judgment. Developmentally speaking (in both a phylogenetic and an ontogenetic sense), functional moral judgment, i.e., judgment motivating behavioral tendencies that regulate social existence, could not have emerged without the involvement of emotional processes. Individual moral judgments, however, can differ widely in the extent to which concurrent emotional activation affects them.

## 5.3 Moral Grammar and the Linguistic Analogy

While Greene and Haidt disagree about the frequency of intuitive judgment and the possibility of socially induced counterintuitive judgment, and possibly focus on different phenomena, they agree that *affective*, intuitive responses play an important causal role in the generation of moral judgments. The linguistic analogy in moral psychology, inspired by theories of John Rawls and Noam Chomsky, disputes this claim. According to the analogy, moral intuitions spring from a small set of innate, unconscious moral principles. The relation between intuitions and principles resembles the relation Chomskyan linguists posit between the sentences we speak and write and an innate, unconscious universal grammar.[700] 'Moral grammar' limits the range of moral codes human beings can adopt – there are moralities we cannot learn. I will not discuss this position in its full breadth, but rather focus on aspects that concern the cognitive architecture of moral judgment.

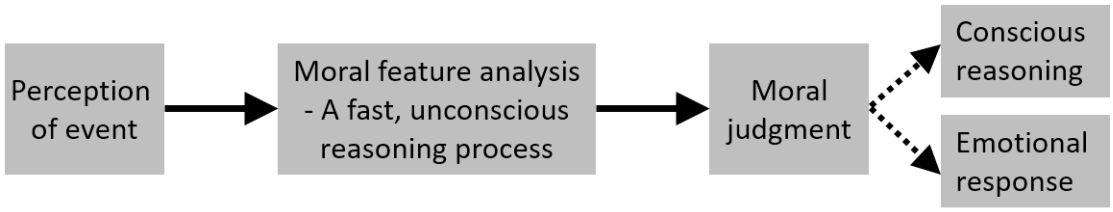


**Figure 6: Hauser's Model of Moral Judgment**

From Nado, J., Kelly, D. R. & Stich, S. (2006): *Moral Judgment (final draft)*. http://www.jennifernado.net/pdfs/papers/NadoKellyStichMoralJudgment2007.pdf (accessed 17.06.2017)

Evolutionary biologist Marc Hauser is a prominent proponent of the linguistic analogy. He disputes that affective responses *determine* moral judgment. On his view, emotions are triggered by *other* mechanisms that identify the morally relevant features of a situation; emotions are therefore *not*, strictly speaking, part of *moral* cognition.[701] He posits a cognitive system referred to as 'moral faculty', concerned specifically with moral matters. According to Hauser, the rules embedded in this moral organ manifest themselves in moral intuitions.[702] The moral faculty performs an unconscious, automatic analysis involving representations of participants, intentions, actions, and outcomes ("who did what to whom and why") and generates judgments based on the results.[703] For instance, judgments about harmful actions

[700] See Dwyer et al. (2010), p. 488. Prinz argues that the moral grammar approach actually constitutes a more strongly nativist view that is usually defended with regard to language. See Prinz (2014), p. 106.
[701] Unlike Haidt, for whom emotions are cognitive processes, Hauser distinguishes between emotion and cognition.
[702] See Churchland (2011), p. 104.
[703] See Dwyer et al. (2010), p. 494, Hauser (2007), p. 45.

are sensitive to intentionality, whether there was direct contact with the victim, or the distinction between actions and omissions.[704] The judgment then determines the emotional response. Hauser and his coauthors believe that the mechanisms that constitute the moral faculty operate largely independent of content.[705] In contrast, Haidt and colleagues emphasize the content-dependent modularity of moral cognition. In their opinion, ascribing all of moral cognition to a single process, module, or moral organ is developmentally implausible: Presumably, different elements of morality evolved as solutions to *distinct* problems of social existence.[706]

In Hauser's model, moral evaluations are largely fixed once the automatic analysis is complete; emotional responses and conscious reasoning do *not* have substantial influence. Situations are presumably analyzed in terms of the basic concepts of moral grammar; distinctions like foreseeing and intending, presence or absence of physical contact, or action vs. omission.[707] The moral grammar approach has been characterized as strongly nativist, that is to say, "the content of our moral values in [*sic*] innately determined or strongly constrained."[708] With respect to the limited influence of conscious reasoning on moral judgment and widespread unawareness of the factors that determine it, the model echoes the intuitionist aspects of Haidt's and Greene's theories. In contrast to those models, however, affect hardly has causal influence on moral judgment in Hauser's model, even though judgment is intuitive.[709] This illustrates that intuition and emotion can be seen as distinct. Hauser's model is an alternative to the traditional juxtaposition of reason and emotion as determinants of moral judgment, and it points to an important feature of moral cognition: The elicitation of emotions has to involve some kind of unconscious analysis of the situation at least sometimes. Otherwise, the regular occurrence of specific, yet rapid emotional responses to certain features of situations is hard to explain. Many psychologists, however, think of these cognitive processes (i.e., information processing) as *parts of* emotions, while Hauser claims that emotions are a separate affair.[710] Can this issue be settled?

Remember the debate about cognitive components of emotions (chapter 4.1). The question was whether emotions *necessarily* contain cognitive elements, so-called appraisals. Non-

704 Sunar (2009), p. 455 mentions that the role of a moral grammar has also been suggested "regarding Fiske's […] relational models and the moral motivations and judgments deriving from them" (section 3.2.3).
705 See Dwyer et al. (2010), p. 497.
706 See Haidt & Bjorklund (2008), pp. 205–206.
707 See Hauser (2007), p. 156, Nado et al. (2009), p. 630, Sunar (2009), p. 455.
708 Prinz (2014), p. 105. See also ibid., p. 106.
709 See Nado et al. (2009), p. 630. However, Hauser et al. (2008), p. 176 state that emotions might also shape the *input* to the judgment system, rather than being mere consequences.
710 See Zimbardo et al. (2006), p. 350.

cognitivist emotion theorists believe they do not. Hauser claims that the moral faculty is distinct from emotional processes; therefore, he does not directly address necessary cognitive *components* of emotions. Yet Hauser's action analysis performs a function very similar to the role appraisals play. While chapter 4.1 dealt with an account that applies to *all* occurrences of emotions, Hauser is concerned only with emotions in the context of moral judgment. It seems uncontroversial that emotions and moral judgment co-occur frequently. We can distinguish two questions regarding this observation: One concerns the *causes* of the co-occurrence. It is possible that emotion causes judgment, that judgment causes emotion, or that a third phenomenon causes both (ignoring, for the time being, the option that judgment *consists* in either emotion or a separate mental phenomenon). A related question is whether emotions that regularly co-occur with moral judgments contain (unconscious) appraisals. Hauser denies that they do and can thus claim that judgments (in the sense of appraisals, not verbal evaluations) cause emotion. However, if these appraisals are construed as *parts* of the emotions, emotions cause judgments (in the sense of conscious verbal expressions of moral evaluations). Note that such an understanding does not require endorsing a *cognitive* theory of emotions in general, or even with respect only to morality. A noncognitive notion of emotions can allow that emotions *sometimes* occur *without* corresponding appraisals and nevertheless hold that emotions frequently involve appraisals, particularly if they occur in the context of social evaluation and moral judgment.

One might think that this is merely an issue of terminology, and that it does not make much of a difference whether appraisals are *parts of* emotions, as long as we agree that appraisals are typically necessary for moral judgment.[711] There are, however, reasons to think that at least in the context of moral judgment, it makes more sense to understand appraisals as parts of emotions: The separation of appraisals and emotions, which in Hauser's account amounts to the exclusion of emotions from moral cognition proper, seems developmentally implausible. Some automatic analysis is required for rapid emotions to correspond regularly to situations with specific features, and Hauser rightly directs attention to this matter. The analytic cognitive ability in isolation, however, does *not* have any effect on fitness. In order to explain how the psychological mechanisms involved in moral cognition evolved, the emotional component is crucial: The motivational function of emotions links appraisals and

[711] "Typically" because some moral judgments, such as the condemnation of apparently neutral behavior reported by Wheatley & Haidt (2005), might occur without appraisals of intentions and outcomes.

behavior, and it is precisely their power to affect behavior on which fitness effects of cognitive abilities depend.[712] Without a connection to fitness effects, it is hard to see why a capacity to appraise what we now consider morally relevant should develop. Awareness of systematic differences in both emotions and the corresponding appraisals across various situations is an important refinement of the understanding of the relation between emotion and moral judgment present in early studies of Haidt and Greene. Not only does emotion play an important role in moral judgment, but different emotions play very distinct roles.[713]

Independently of the involvement of emotions in moral cognition, Hauser's claim that some moral codes could not take hold in human societies because they are incompatible with the grammar of the moral organ is contentious. Philosopher Patricia Churchland mentions human sacrifice, variation in the treatment of women across epochs and cultures, and the atrocities committed in the dictatorships of the twentieth century to point out that the alleged moral grammar does not even prevent establishment of extremely inhumane norms.[714] This criticism hardly seems decisive. As long as we do not know what the principles of moral grammar consist in, it is hard to argue that any observed behavior contradicts these principles. Similarly, norm variation across moral systems conflicts with the linguistic analogy *only* if we assume that the variation concerns the evaluation of features of iudicanda which we would expect to figure in the principles of moral grammar in a specific way. Since these principles are as of yet unknown, variation in itself is not evidence against the theory. Churchland also argues that Hauser's model does not cover many instances of moral judgment, since it leaves even less room for causal influence of conscious reasoning than Haidt's SIM. "[The model] may apply in some situations such as seeing a child choking at dinner, but it clearly does not apply in multitudes of other situations, such as whether to go to war against a neighboring country."[715] It is true that Hauser and his coauthors focus on fast, unconscious, and automatic action analysis. Connecting this morality-specific debate to the more general debate about the cognitive components of emotions, however, reminds us that appraisals are not necessarily generated exclusively by system 1 mechanisms. In some cases, it certainly takes controlled reasoning to understand an agent's intentions and knowledge, action consequences, and other relevant features of a iudicandum.[716] From the

712 The argument depends on the assumption that, at least in a significant number of morality-related cases, motivation is tied to emotion. If motivation regularly occurred without emotion, the implication of emotion would not be as straightforward. However, since motivation is typically understood as part of emotions in psychology, the assumption seems justified.
713 See Chapman & Anderson (2011), p. 255.
714 See Churchland (2011), p. 106.
715 Ibid., p. 111.
716 See Chapman & Anderson (2011), p. 257.

perspective of the linguistic analogy, one could still argue that these 'reasoned' appraisals precede emotions, and that controlled cognition merely alters *the input* fed into the intuitive principles of the moral organ, not the principles themselves.

In sum, moral grammar theory is not decisively disqualified by arguments that it is incompatible with observable inhumane norms, or that it cannot (at least in modified form) account for a causal role of conscious reasoning. However, the incompatibility charge does raise a significant challenge for proponents of innate moral grammar: If their defense rests on the claim that we do not yet know what the principles of the moral organ are, the claim that there are such principles is weakened considerably. Despite these problems, moral grammar theory rightly directs attention to the automaticity and regularity of emotional responses to specific iudicanda, and the hypothesis that this regularity is the expression of *innate* appraisal-response couplings (the moral grammar) remains respectable.

## 5.4 Emotion and Moral Judgment

What do these models imply regarding demands for less intuitive or emotional, more 'reasoned' moral judgment? Empirical evidence is still too scarce to allow detailed conclusions regarding the sequence of events involved in moral judgments. Current imaging technologies do not offer the kind of temporal and spatial resolution necessary to support one or the other model; in fact, more are conceivable. Imaging data indicate only that brain areas *associated* with some mental phenomenon (emotional processes, short-term memory, etc.) are more or less active at specific time intervals. It is possible that these areas actually contain several neural substructures: Because of the limits in spatial resolution, we cannot 'see' whether hypothetical neural substructure A or B is active, we only know that the area containing A *and* B is active. If in fact A is active while only B corresponds to some specific process, conclusions about the psychological processes involved in a behavior may be mistaken. Similar caveats regarding temporal resolution are in order: Whether activation of emotion-related areas precedes activation of the prefrontal cortex (for instance) or vice versa, or both happen simultaneously; every variant would presumably look the same on images because transmission of electrical impulses from one part of the brain to another is very fast.[717] Furthermore, little is known about how social knowledge and other concepts relevant to moral judgment are represented in neural activation patterns, and about how these activation patterns relate to the subjective experiences involved in moral judgment,

[717] See Huebner et al. (2009) and Hauser et al. (2008) for overviews of questions regarding the details of moral judgment which could not, at that time, be answered on the basis of available data.

i.e., what a given mental event 'feels like'. For all we know, Hauser's action analysis might *consist in* combinations of activations which represent the relevant categories of social cognition and activations of the limbic system associated with emotional mechanisms and other functional modules. Particular combinations of activations might 'feel' like a specific moral emotion.[718] Moreover, the three models sketched are not mutually exclusive. In fact, there appear to be various alternative mechanisms that generate moral judgments.

### *5.4.1 Emotion and the Point of Decision*

We have now considered two models of moral judgment (Haidt's social-intuitionist model and Greene's dual-process model) that emphasize the role of affective intuition (system 1) in moral judgment generally, but differ in their assessment of the possibility and frequency of judgments shaped by system 2 processing. A further model (the linguistic analogy) attributes the 'core' of moral judgment entirely to a *nonemotional* subdivision of system 1. Above, I have argued that individual moral judgments presumably involve reasoning and emotional-intuitive processes to different extents, and that a debate about a *single* correct model with respect to these variables might betray an inappropriate desire to simplify. Nevertheless, we can still ask whether the available empirical evidence regarding the stage or 'point' at which judgment is fixed speaks in favor of one of these models, and whether this point can be located in distinctions between emotions, intuitions, and controlled processing at all.[719] Greene et al.'s early neuroimaging results indicated that certain types of moral judgment correlate with increased emotional activation, others with elevated activity in areas associated with controlled processing. Yet, due to the limited temporal resolution of fMRI, it remained unclear whether emotions and reasoning *cause* judgments, or whether they are simultaneous or minimally delayed accompanying phenomena of moral judgment proper, as in Hauser's proposal. Experiments with disgust elicitors such as dirty desks and fart sprays showed that, at least sometimes, changes in emotional activation generate changes in moral judgment, making condemnation more severe.[720] A study in which funny video clips increased the frequency of consequentialist judgment in moral dilemmas provides complementary evidence: Differences in emotional activation can not only intensify a judgment, but also make the evaluation switch from impermissible to permissible.[721] The hypnosis study by Wheatley and Haidt showed that emotions can generate condemnation of

[718] See Moll et al. (2005) for hypotheses regarding the role of various brain areas in moral cognition.
[719] The discussion in the next paragraphs is based on Prinz (2012), pp. 297–302.
[720] See Schnall et al. (2008).
[721] See Valdesolo & DeSteno (2006).

actions that otherwise appear neutral. These results speak against the moral-grammar account, since situation descriptions, and thus the corresponding appraisals presumably do not change. They also appear to count against positions according to which moral judgment depends mainly on reasoning. The results are compatible, in principle, with the view that appraisals or controlled processes are necessary for moral judgment, and that emotional stimuli take effect by altering these. It remains unclear, however, *how exactly* reasoning is supposed to differ in these cases; moreover, since judgments change even though *none* of the objects of automatic analysis (agents, intentions, consequences, etc.) proposed by Hauser seems to vary, it is tempting to argue that emotions determine judgment *at least in addition to* 'moral grammar'. However, some studies indicate that the moral evaluation of an act affects attributions of intentionality (the so-called Knobe effect or side-effect effect).[722] If such surprising feedback from affective or evaluative processing to the appraisals that the moral faculty processes are common, a suitably modified version of the moral-grammar hypothesis might still be viable. Such a modification would constitute a departure from the original moral-grammar approach insofar as it allows that emotions have a *causal* effect on judgment. If we assume that such feedback mechanisms are uncommon, then it seems that emotions affect moral judgment *even if* the factors assessed by Hauser's moral faculty remain fixed. It is conceivable that *other* appraisals are affected (for instance, an appraisal to the effect that disgusting smell indicates the presence of an offensive substance), but it is also conceivable that the elicitation of disgust through hypnosis or the presence of core-disgust elicitors such as noxious smells resemble the elicitation of emotions through music, and do not involve appraisals.[723] Possibly, specific emotions that affect moral judgment can be triggered in various ways, including, but not limited to those that made these emotions evolutionarily advantageous. Maybe only some of these ways involve appraisals, while more direct manipulation of the corresponding physiological processes is also possible. This idea provides an explanation of the behavioral results mentioned above: Influences without causal relation to the iudicandum, such as odors or hypnosis, can affect emotions. It also allows that the set of triggers that elicit emotion-specific appraisals expands in environments that differ from the EEA.

Haidt's observations of dumbfounding provide further support for the view that affect can be sufficient for (maintaining) moral judgment: Subjects sometimes hold on to their evaluations even though all their justifications have been defused. Then again, other studies

[722] See Knobe (2003).
[723] See Chapman & Anderson (2011), p. 256 for a multiple-appraisal model of moral emotions.

show that the inability to articulate the principles underlying judgment differs across principles: In experiments designed to test sensibility to the distinction between *actions* and *omissions*, judgments are typically in line with the 'action principle' (harmful *acts* are worse than harmful *omissions*), and subjects are usually able to name that difference. If they generate a pattern of judgment according to which harm as a *side effect* is not as bad as harm used as a *means to an end* (intention principle), however, they are frequently unable to articulate it (dumbfounded).[724] While these findings do not prove that the principles people articulate also *cause* their judgment, and leave unanswered the question which system was dominant in the generation of both judgment and principles, they fit the view that morality, and moral judgment more specifically, involve many constellations of mental mechanisms. Thus, generalizations about the nature of moral judgment and the 'point of decision' should be advanced only with great caution. Rather than show that emotions are decisive for 'moral judgment in general', I take it that the available evidence *allows* that emotions are the decisive factor in *some* judgments plausibly labeled 'moral'.

### *5.4.2 Is Moral Judgment Without Emotion Possible?*

The choice of words above already suggests that emotions might be involved, but not dominant, in other moral judgments. In order to map the terrain of moral cognition, we can consider an extreme case and ask whether emotions are *necessary* for moral judgment. While we should be cautious not to overextend the range of empirical findings, research on psychopaths provides a strong case in point: Simplified characterizations portray them as fully rational, but capable only of generally flattened affect. If ordinary moral judgment did *not* require emotions, moral judgments of psychopaths should not differ from those of non-pathological individuals. Yet, psychopaths do not make normal moral judgments (chapter 4.4.3). In particular, they do not draw the distinction between conventional and moral rules common to Western notions of morality. They tend to treat all violations as *moral* violations in terms of response patterns (serious, judgment universally valid), possibly, if incarcerated, in order to demonstrate that they are fit to be released. On the other hand, psychopathic individuals seem less strongly motivated than nonpsychopathic individuals are to *act* according to these judgments. Presumably, psychopaths see both moral and conventional norms as rather arbitrary and obey them only if it serves their own interest. These observations prompt the question whether the judgments psychopaths make should count as *moral* judg-

[724] See Cushman et al. (2006), pp. 1086–1087.

ments at all. This puzzle brings into focus a problem that affects many psychological models: The definition of 'moral judgment' is often vague or implicit. This vagueness is due to the use of the term in ordinary language, and I believe it is legitimate to operate with this fuzzy concept. Acknowledging that multiple psychological phenomena fall under the heading of 'moral judgment' makes it easier to see various models as being complementary rather than in opposition.

Even on such a liberal approach, psychology offers new perspectives on debates about the necessary and sufficient conditions that characterize moral judgments, such as the debate about judgment internalism and externalism in metaethics. While internalists hold that *genuine* moral judgments automatically motivate to act accordingly, externalists deny that congruent motivation is a necessary component of moral evaluations. On an externalist account, the judgments psychopaths make about moral issues might count as moral judgments. With a view to phylogenesis, however, the internalist account seems more convincing. Firstly, the reasons that make Hauser's claims about appraisals more convincing if appraisals are understood *in conjunction with* emotions might also be brought forward against the externalist position: It is hard to see the evolutionary advantageousness of a judgmental capacity that lacks the action-guiding force provided by emotional mechanisms. The externalist might respond that the advantageousness of a capacity to perceive how others evaluate actions and act accordingly if it serves one's own interest is not at all hard to see. Possibly, such a strategy was just too risky and therefore crowded out by immediate, nonstrategic motivation that generates behavioral norm conformism. The explanatory force of the emotional-developmental story provides a stronger argument for phylogenetic (possibly also ontogenetic) moral judgment internalism: While a capacity to perceive evaluations made by others is advantageous also if the motivation to conform is self-interested, it is hard to explain the regularity of those perceived evaluations on this account. Judgments made by psychopaths are parasitic on the emotional capacities shared by the nonpsychopathic majority. The evaluations psychopaths imitate would look very different, had emotion-elicitor couplings, which include a motivational aspect, not shaped our notions of what is and is not morally relevant. In sum, it seems that emotions are sufficient, necessary, or both for at least some moral judgments, and certainly necessary for the development of morality and moral judgment as we know it.

Prinz discusses three objections to or concerns about the view that moral evaluations *are* emotions, also known as 'sentimentalism'.[725] The first is that even if emotions strongly affect

[725] See Prinz (2012), pp. 302–304.

moral judgments *in practice*, they *should not*. Alas, a challenge David Hume posed to rationalists is still relevant: Without emotions, how do we know what we value? If we agree that motivation requires emotion, then 'pure reason' will get us nowhere; the call for purely rational moral judgments disregards human psychology. The second objection concerns the descriptive adequacy of accounts of morality that emphasize emotion. Does not the ubiquitousness of *deliberation and debate* about moral issues prove that reasoning is central to moral judgment? Haidt's social-intuitionist model (chapter 5.1) and the research it is based on indicate that a significant percentage of this reasoning is in fact rationalization; the 'press secretary' at work. What about the rest? I do believe that reason plays an important role in constructing analogies, comparisons, and other cognitive devices that can affect moral judgment. Moreover, it is important to indicate precisely to which phenomena one is referring. It might be the case, as Greene has suggested, that while much of our *everyday* moral judgment might be 'unsophisticated' and emotion-driven, moral *progress* comes from effortful applications of the rational faculties. New conclusions spread and become habitual, emotional, and intuitive for generations to follow. His analogy is technology: While technology surrounds most inhabitants of industrialized countries, hardly anyone makes a technological invention.[726] A sentimentalist might concede some truth to this view, but argue that those who initiate or catalyze moral change nevertheless would not know what to value, were it not for their emotions. Widespread adoption of changes in moral views seems similarly inexplicable without reference to the emotional appeal of a newly developed or adjusted position. In addition, it might be useful to imagine a continuum from *basic* to *derived* norms and values.[727] While a (conditional) prohibition of harming others is rather basic, other norms require elaborate argument in order to relate them to these basic values. Thomas Pogge's argument that the present economic world order other systematically disadvantages inhabitants of developing countries and thereby causes severe harm is an example of such an effort.[728] Derived norms about the proper design of the relevant national and international institutions are moral, but hard to explain without significant rational input.

A third way to criticize sentimentalism is to cite cases in which individuals apparently make moral judgments *in spite of* their emotions, like those subjects in Haidt's studies on harmless transgressions who judge that actions are gross, but *not wrong*. In defense of sentimentalism, Prinz proposes a model of competing emotions: While people consider harm,

[726] A comparison Greene made at the conference "The Moral Brain", New York University, March 2012.
[727] See ibid., p. 303.
[728] See Pogge (2005).

or rather its absence, *more* relevant than disgust in these cases, the presence of harm nevertheless is relevant only because of its *emotional* effects. Judgments in which emotional responses are overridden by *other* emotions, or by habitual evaluations whose existence depends on emotional responses, do *not* show that moral judgment is not ultimately emotion-driven.[729]

## 5.5 Generating New Intuitions: Implicit and Explicit Processes

In the preceding chapters, particularly in the description of the origins and functions of emotions, evolutionary processes figured as primary origin of intuitive-emotional responses in moral judgments. The following chapter presents and discusses other hypotheses about the origins and modifications of intuitions. Some of these were proposed as rationalist objections or amendments to Haidt's social-intuitionist model and Greene's dual-process model, both of which (the social-intuitionist model slightly more so, in my impression) are often perceived as granting too much importance to system 1- relative to system 2 processes. While many judgments are made rapidly, without much conscious moral reasoning, slow and careful judgment processes also occur. Sometimes, it takes hours of deliberation to reach a decision. These observations indicate that there are indeed several ways to judge morally, each of which might employ a slightly different combination or succession of mental mechanisms. In the eclectic spirit of this thesis, I integrate various models of intuition change in a broad picture of (different sorts of) moral cognition, and argue that none of the models undermines the thesis that emotional intuitions fundamentally shape morality.

### *5.5.1 Explicit Processes: Appraisal Shifts and Input Selection*

Psychologists David Pizarro and Paul Bloom argue that the social-intuitionist model fails to capture important ways in which reason can alter intuitions and thereby affect moral judgment. In the SIM, all nonsocial modification of judgment takes place either via 'reasoned judgment' (independently of intuitions; link 5 in Figure 4) or private reflection (which generates new intuitions that affect judgment; link 6 in Figure 4). Pizarro and Bloom suggest that deliberate modifications of the *appraisals* that generate intuitive-emotional responses and the deliberate selection of the intuition-eliciting *situations* one confronts constitute additional ways for reasoning to determine the output of intuitive processes.[730] Deliberate

[729] See Prinz (2012), p. 304.
[730] See Pizarro & Bloom (2003), p. 194.

appraisal changes modify the interpretation of specific inputs: For instance, empathic responses can be elicited by instructing people to take the perspective of others (see also chapter 4.4.3). The reasoned modification of intuitions can even override self- or group-interest. In such cases, reasoning, rather than evolved intuitions, appears to determine judgments. We can also willingly modify our implicit attitudes by putting ourselves into situations we expect to change them.[731] While Pizarro and Bloom believe that their model is compatible with the mechanisms of the SIM, they claim to attribute more importance to deliberative reasoning. People sometimes take a moral stand that is *not* in line with either their socialization or their evolutionary interest, and the social-intuitionist model does not appear to capture this possibility adequately due to its emphasis on evolutionary and social determination of intuitions and judgment. We often face moral decisions about which we have no basic intuitions, such as determining the proper balance of family and work, giving to charity, etc. While judgments affecting such issues are often automatic, it is conceivable that their *origin* lies in careful reasoning, the results of which have later become habitual responses.[732]

Bloom is particularly interested in processes that might count as moral progress. While many of these, in particular those which fit under the heading of an expanding circle of moral concern, can be explained by extensions of contact and interdependence with other people and without reference to deliberative processes, others cannot. In his view, the interplay of reason and emotions sometimes generates important new moral insights:

> This process is similar to what goes on when we generate other sorts of ideas, including philosophical and scientific ones. As a core example of this, Singer (1981) argued that the great insight about morality is the notion that it should be built from an objective position. Put crudely, the idea here is that nobody is special, which is an insight enshrined in the Golden Rule, the "impartial spectator" of Adam Smith, and the "original position" of John Rawls. We have the capacity to generate such ideas, and they really do matter, shaping the societies in which we live.[733]

Haidt believes that the social-intuitionist model can incorporate these effects of prior reasoning on intuitions involved in solving real-world moral problems.[734] While he agrees that new appraisals can engender new intuitions, Haidt points out that these *new* appraisals are more likely to stem from social interaction than from private reflection. Similarly, we can

731 See ibid., p. 195.
732 See ibid.
733 Bloom (2012), p. 85.
734 See Haidt (2003b), p. 197.

choose to make new acquaintances and expect our intuitions to become more like theirs over time. A slight disagreement remains between Haidt on the one hand side, and Pizarro and Bloom on the other not regarding the possibility, but the *frequency* of these processes. According to Haidt, both deliberate *private* generation of new appraisals and deliberate exposition to environments that foster counterintuitive attitudes are infrequent, thus, reasoning in moral judgment should not be portrayed as the 'dog wagging an emotional-intuitive tail', but vice versa.

Pizarro and Bloom's description of mechanisms by which effortful deliberation *modifies* intuitions is a valuable contribution. While attention to evolutionary influences is crucial for an understanding of the foundational function of the intuitions involved in moral judgments and the kinds of concerns they establish, the extent of cultural and individual variation on these themes should not be underestimated.

### *5.5.2 New Intuitions from Implicit Learning*

In addition to what Pizarro and Bloom suggested, new, culture-specific intuitions also emerge *without* prior reasoning. Socialization can anchor powerful intuitive judgments in individual minds, for instance in southern states of the U.S. during the first half of the twentieth century:

> [A] Black woman and a White man decide to get married; […] a Black boy who is 15 years old drinks from a water fountain designated 'for Whites only.' […] [L]arge numbers of White people would have had strong gut reactions that all these acts are wrong. They would have maintained that they know they are just wrong.[735]

Here, system 1 processes *learned* rather than innate judgments. Learning does not necessarily involve system 2 processing regarding the *content* of the corresponding norms. Note that while 'intuitive' moral judgments occur not only in response to elicitors present in the EEA, learned triggers can nevertheless build on evolved features of the psyche, such as, in the examples just mentioned, a concern for group belongingness, possibly coupled with disgust felt towards out-groups. Learned concepts can be applied intuitively if they are well understood and their application to a given situation is straightforward.[736]

---

[735] Turiel (2006b), p. 19. In Turiel's view, this is a counterexample to Haidt's intuitionist model, since *intuitive* judgments supposedly cannot involve *complex* appraisals concerning group-membership, social relationships, and perspectives on society. According to the criteria employed in this thesis, however, intuitiveness refers not to content, but to the way in which judgments emerge: Intuitions are quick, effortless, and not preceded by conscious weighing of reasons. Turiel's example satisfies these requirements.

[736] See Turiel (2006a), p. 819. Again, Turiel would not refer to such judgments as intuitive, but with regard to the distinction between systems 1 and 2 outlined above, they are.

James Woodward and John Allman, who work in philosophy and neuroscience respectively, emphasize that the implicit learning that generates intuitions in ontogenesis warrants more optimism regarding their reliability than many of their colleagues seem to have. They argue that intuitive influences should not be eliminated from all decision making in the moral domain. Intuitions, in combination with system 2 processing, can produce adequate judgment if their ontogenetic development satisfies certain conditions (it remains uncontroversial that intuitions are sometimes biased).[737] They rely on Haidt's definition of moral intuitions (sudden appearance in consciousness of an evaluative feeling, see chapter 5.1) in a dual-process framework that distinguishes between fast, automatic, largely unconscious processes and conscious, effortful, and slow deliberative processes.[738] Based on neuroscientific evidence, they hold that emotions are important for intuitive, automatic responses in moral judgment. In contrast to researchers like Haidt or de Waal, they assume that the emotional neurobiological systems involved in moral intuition are evolutionarily recent and unique to humans, while other primates show at most similar, but more primitive capacities.[739] Woodward and Allman argue that in nonmoral cases (prudential or personal decisions), unconscious emotional processing and intuitions can generate better decisions than conscious reasoning, and that something similar might be true for moral judgments. This is because implicit learning from environmental feedback to decisions made by the self or others can shape intuitions.[740] Note that Woodward and Allman focus on *ontogenetic* development, rather than innate features of intuitions shaped by supraindividual 'learning' through natural selection. Implicit learning processes capture the overall effects of complex decisions more accurately than reasoning, because cognitive limitations prevent the conscious, reasoned consideration of *all* relevant decision aspects. Specifically, conscious application of *moral* decision rules often neglects the mental states (intentions, desires, beliefs, etc.) of interaction partners because these states can only be assessed via emotional simulation, rendering such rules unsuitable for the strategic nature of decision making in social contexts. This deficit is particularly pronounced in 'parametric' variants of utilitarianism that reduce the relevant dimensions of a situation to the numbers of lives lost and saved or similar observables.[741] 'Strategic' utilitarianism, in contrast, takes into account motives and intentions, future incentive effects, etc. If learning from corrective feedback is possible and

737 See Allman & Woodward (2008), p. 172.
738 See ibid., p. 167.
739 See Woodward & Allman (2007), pp. 188–190.
740 See Allman & Woodward (2008), p. 169.
741 See Woodward & Allman (2007), p. 200.

tracks features that are uncontroversially relevant for moral evaluation, moral intuitions should be taken seriously even from a (strategic) utilitarian viewpoint.[742] Mental states matter from the more comprehensive strategically utilitarian perspective, and intuitions shaped by implicit learning may be the best option to factor in mental state information. Intuitions do not have to incorporate personal experience to be potentially trustworthy; they can be based on mental simulation or experiences made by others. Moreover, we can learn by focusing on similarities between the new situation being evaluated and familiar situations.[743] Under these conditions, emotional responses can guide decisions well even if we are unaware of the principles underlying these responses. Contrary to what parametric utilitarianism suggests, absence of 'rational reconstructions' of an intuition's origin does not necessarily invalidate that intuition.[744]

Decision making based on emotional intuitions is most likely to be advantageous in high dimensional, complex decision problems, and moral decisions, involving the interaction of human beings, are frequently of that kind. Since real-life decisions are often too complex to process consciously in real time, Woodward and Allman are skeptical of (thought-) experimental vignettes stripped of many details or otherwise unlike what people actually encounter.[745] Intuitive decisions might be suboptimal under these artificial conditions, but outperform conscious deliberation in the real environments to which implicit learning processes have tailored them. Intuitions are apt to track morally relevant properties since they (unconsciously) evaluate the social consequences of our decisions: how others will respond in realistic scenarios.[746]

According to Woodward and Allman, the association between emotional processing and deontological judgments reported by Greene might be due to the fact that emotional processing is required for proper assessment of mental states, and the fact that mental states such as intentions, or concepts that involve empathy (such as dignity and respect), are important in deontological ethics. These considerations apply to complex problems in particular; simple problems sometimes allow for the application of deontological rules ("Never lie!") that do not require emotional processing.[747]

> We would expect emotional processing to be particularly likely to be involved when the choice is complex and high dimensional, where there is no consciously accessible

[742] See ibid., pp. 37–38.
[743] See ibid., p. 38.
[744] See ibid., p. 24.
[745] See ibid.
[746] See ibid., pp. 32–33.
[747] See ibid., pp. 31–32.

> rule indicating what to do, and where emotional processing can play [an] integrating and synthesizing role […].[748]

The potential superiority of unconscious processing with respect to complex problems does not imply that is *generally* better to make decisions based on incomplete information or little experience; rather, emotional intuitions shaped by implicit learning are most likely to produce good choices if they operate on a large body of information.[749] The authors moreover acknowledge that emotional processes can be subject to biases and framing effects.[750] Yet, Woodward and Allman believe that good moral decision making "requires the integrated deployment of both the automatic and deliberative systems (and cognition and emotion) working together and mutually supporting one another."[751]

One might object to this optimism regarding intuitive decisions. In nonmoral cases, what counts as a good solution to a problem and which features of a situation intuitive processes should track is often uncontroversial. In moral decisions, in contrast, the right choice and, relatedly, the *relevance* of iudicanda features can be quite controversial. (Recall the trolley dilemma.) Thus, firstly, the criteria for what constitutes a good decision might be substantially different in the moral case, such that superiority in nonmoral cases does not imply or make more probable the superiority of the same process in moral judgment. Secondly, the contentious character of the criteria might not allow for an uncontroversial evaluation of the output produced by intuitive processes.[752] In response, Woodward and Allman argue that if empirical evidence showed that moral judgment without *any* intuitive processing neglects *uncontroversially* relevant factors, and that those who lack intuitive capacities tend to make judgments *most* moral theories consider inadequate, then 1) *some* moral intuitions respond to morally relevant factors and 2) they may even be *necessary* to track these factors. Whether a *specific* intuition tracks morally relevant factors is an empirical question.[753] Accordingly, empirical research, in combination with information about what is uncontroversially relevant, could identify trustworthy and unreliable intuitions.

Woodward and Allman claim that the emotional intuitions shaping moral judgment are evolutionarily recent and subject to significant ontogenetic modification and amendment

748 Ibid., p. 32.
749 See ibid., p. 30.
750 See ibid., p. 7.
751 Ibid., p. 19.
752 See Kauppinen (2014), p. 295 for a similar argument: He claims that the development of intuitive expertise requires, firstly, regularities in the environment in which the intuition develops, and secondly, "rapid and unequivocal feedback". Kauppinen doubts that the second condition is met because feedback consists mostly in the opinions of others, whose relation to 'moral truth' is unclear.
753 See Woodward & Allman (2007), pp. 193–194.

through implicit learning. How does this view integrate with the models presented so far? In the interest of a nuanced understanding of moral cognition, it is important not to neglect the influence of ontogenetic development. Evolution does *not* fully determine moral intuitions; we can develop all kinds of intuitive valuations. My suggestion here is, rather, that intuitions attuned through implicit learning to stimuli *not* related to fundamental moral concerns are infrequent or temporary. Even if, as Woodward and Allman claim, specific emotional processes in moral judgment are unique to humans and evolutionarily recent, they are still shaped by evolutionary processes our ancestors underwent. In fact, evidence of the sort presented, for instance, in chapters 4 and 6.2.3, indicates that at least protoversions of many of these capacities are present in our evolutionary relatives. Evolved learning modules of the kind suggested by Haidt and Joseph can accommodate the ontogenetic changes emphasized by Woodward and Allman.

### *5.5.3 The Diachronic Penetrability of Moral Intuitions*

Richmond Campbell and Victor Kumar present an empirically supported explanation of moral change which contains yet another model of intuition change based on 'moral consistency reasoning', which differs from the options of conscious appraisal modification and deliberate input control suggested by Pizarro and Bloom at least in its focus on the role of consistency.[754]

> "Moral consistency reasoning," [MCR] as we call it, exposes latent moral inconsistencies, embodied in conflicting moral judgments about cases that are, by one's own lights, similar in morally relevant respects.[755]

Campbell and Kumar's minimalist-moral-dual-process model (MMDP) attributes most of the usual characteristics to system 1 and system 2, but is silent on whether system 1 and system 2 are innate or learned and associative or rule-governed. Intuitive moral judgments rest on moral *norms* to which we frequently do not have conscious access.[756] Importantly, they characterize system 1 as impenetrable by or encapsulated from *simultaneous* system 2 processing: System 2 does not affect concurrent system 1 output. Over longer periods, however, system 2 processing can affect system 1 processing; system 1 is synchronically impenetrable but diachronically penetrable. The diachronic penetrability of system 1 is the

754 See Pizarro & Bloom (2003).
755 Campbell & Kumar (2012), p. 274.
756 System 1: domain specific, affective, fast, automatic, effortless, impenetrable, unconscious; system 2: domain general, cognitive, slow, controlled, effortful, penetrable, conscious. See ibid., p. 277.

basis of substantial moral change, since, in MCR, system 2 and system 1 together produce long-term changes in emotional responses.[757]

What about the penetrability of system 1? According to a widely held view, system 2 processing does not affect visual perception (part of system 1). For instance, the Müller-Lyer illusion is unaffected by knowledge that both lines are of equal length. With regard to moral intuitions, authors like Woodward and Allman or Bloom claim that such system 1 outputs *can* change, and that existing dual-process models fail to capture this possibility adequately.[758] While Woodward and Allman emphasize mechanisms internal to system 1 (implicit learning), Bloom is concerned with "rational deliberation and debate"[759], i.e., system 2 processing. Campbell and Kumar agree that moral intuitions can change over time and focus on the interaction between system 1- and system 2 processing leading to such change. In their view, Greene and Haidt neglect long-term effects of moral reasoning on emotional intuitions that account for significant moral change (e.g., the abolishment of sexist and racist attitudes, and of the corresponding emotional intuitions).[760]

> In […] consistency reasoning, one response is independently less tenable or easier to relinquish than the other because of already established patterns of moral response, and unless a relevant difference between the two instances is found, a person engaged in this reasoning must either accept inconsistency and thereby give up any semblance of having a good reason for either of the responses or else revise the less tenable response.[761]

If, for instance, conscious comparison of heterosexual and homosexual relationships reveals no morally relevant differences, the resulting, discomforting perception of inconsistency, after a while, entails change in intuitive evaluations of homosexuality.[762] While Greene and Haidt allegedly propagate a false emotion/reason dichotomy, moral consistency reasoning models the interaction of both systems more adequately, with a focus on reasoning.[763] Expert chess players, for instance, can assess constellations intuitively, even though *acquisition* of the principles underlying these assessments requires prototypical system 2 processing.

---

757 See ibid., p. 275.
758 See ibid., pp. 282–283.
759 Bloom (2010), p. 490.
760 See Campbell & Kumar (2012), p. 286.
761 Ibid., p. 284.
762 The change may remain incomplete; see ibid., p. 287. Campbell and Kumar conceive of this response as importantly different from dissonance reduction as typically discussed in psychology, since, in the moral case, individuals are *aware* of the dissonance between their beliefs. See ibid., p. 297, note 51.
763 See ibid., p. 283. Moral consistency reasoning is called a mechanism of "reason-based moral change" on pp. 275 and 283.

Campbell and Kumar's characterization of the positions of both Greene and Haidt, however, seems slightly off target. Haidt's social-intuitionist model distinguishes different effects of reasoning, as opposed to the mere *presence* of other opinions. Social reasoning, represented by the reasoned persuasion link in Figure 4, can certainly lead to substantial changes in moral intuitions, but it does so mainly because arguments trigger other or new intuitions. Less frequently, change results from private reflection or even 'reasoned judgment'. Based on the work of Greene and colleagues, I have argued in chapter 5.2 that the social-intuitionist model should be amended by the possibility to arrive also at *counter*intuitive judgments in response to socially presented arguments. Campbell and Kumar are not concerned with counterintuitive judgments. How does moral consistency reasoning relate to the SIM? According to Campbell and Kumar, all effects of *social* reasoning in Haidt's model rest on motives of conformity.[764] That is incorrect: While the *social* persuasion link indeed represents the effect of conformity motives (the mere *presence* of opinions expressed by esteemed or likeable individuals generates corresponding intuitions), the *reasoned* persuasion link represents the influence of *arguments* presented by others (the *content* of the argument triggers intuitions).[765] This understanding of intuition as a bottleneck constitutes a fundamental difference between Campbell and Kumar's approach and Haidt's model. Campbell and Kumar claim that both Greene and Haidt "are silent about the possibility of reasoning influencing the operation of the intuitive system."[766] My impression is that Haidt has the processes described by Campbell and Kumar plainly in view, and would argue that intuitions, rather than reasoning, drive attitude change. Haidt could reconstruct (social) moral consistency reasoning as an instance of 'reasoned persuasion': Amy has an intuition $F_X$ regarding some subject matter X. Bob, however, is of the opinion that $G_X$. According to the SIM, Bob can convince Amy by triggering intuition $G_X$ in her mind. How can Bob do that? One way is to *argue* that X is in fact like Z, on which both Amy and Bob share intuition G. For Amy to accept that G is the adequate judgment also regarding X, she has to be unaware of any morally relevant differences between Z and X, or at least of any differences substantial enough to undermine judgment $G_X$. The point is that, according to Campbell and Kumar's own model, Bob would have no leverage on Amy's opinion if he could not recruit their strong, shared *intuition* regarding Z. Convincing Amy that there are no relevant

764 See ibid., p. 288.
765 See Haidt & Bjorklund (2008), p. 191.
766 Campbell & Kumar (2012), p. 288.

differences between X and Z is one way in which *arguments* can trigger the (contextually) new intuition G in response to X.

Campbell and Kumar might point to *nonsocial* instances of consistency reasoning; moral exemplars who resolve inconsistencies of which their contemporaries are still unaware or in which they temporarily persist.[767] Such cases are similarly unproblematic for the SIM: Firstly, it will be hard to show that these exemplars experience 'inconsistencies' in their moral attitudes *independently* of social triggers or new intuitive responses. If a slaveholder comes to believe that slavery is immoral, this new view will likely spring from intuitions triggered either by new or changed perceptions (new trigger/private reflection) or by a vivid portrayal (reasoned persuasion) of the harm done to the slaves. Secondly, examples of persistent *counter*intuitive attitude change in which *neither* of these processes occurs would make a strong case in favor of a decisive role for reasoning, but would not contradict the SIM. In the SIM, these examples could count as rare instances of reasoned judgment. The fact that achieving something of the sort counts as exceptional supports this frequency estimate. Moreover, the 'new' attitudes Campbell and Kumar are interested in form only a subset of the changed attitudes conceivable in the SIM framework: Moral consistency reasoning retains some intuitions. In a sense, then, the social-intuitionist model leaves *more* room for effects of reasoning than the MCR model.

Is reasoning being neglected in my Haidtian reformulation of moral consistency reasoning? Think about the operations reason performs in Campbell and Kumar's model. Reason recognizes a similarity between situations (X and Z) that elicit conflicting judgments, and it identifies differences between X and Z. This is certainly a precondition for intuitive responses to these differences. Yet claiming that this role for system 2 processing is a substantive correction of Haidt and Greene seems exaggerated. In a further step in Campbell and Kumar's model, reasoning, triggered by an unpleasant perception of inconsistency, attempts to revise the less tenable responses.[768] Reasoning, however, does not appear to be the driving force behind the attitude change even on Campbell and Kumar's own account. The initial responses are *intuitive*, the moral relevance of the differences and thus the presence of an inconsistency is assessed *intuitively*, and an *intuitive* response to inconsistency provides the motivation to revise the original intuitions. To my mind, these are good reasons to emphasize intuition, even though reasoning certainly provides some crucial input to the intuitive systems.

767 See ibid., p. 289.
768 See ibid., p. 293.

Apart from the uncharitable characterization of Haidt's position, Campbell and Kumar's close look at moral cognition nevertheless has merit: They spell out in more detail some effects of reasoning to which Haidt only alluded. Moreover, they explain how reasoning can figure in episodes of moral change that occur via *elimination or weakening* of certain intuitions. This option is consistent with, but not emphasized in the SIM, which concentrates on the *elicitation* of intuitions. Moral consistency reasoning starts from at least two opposing emotional system 1 responses to iudicanda. System 2 notices similarities between the cases, identifies differences, and feeds them back to system 1. System 1 checks for moral relevance, and, if none of the differences is (sufficiently) relevant, an unpleasant system 1 response to dissonance motivates abolishment of that initial response which is less firmly anchored in other patterns of moral response. Within the SIM, such a reevaluation of a iudicandum X is based on the *elicitation* of a new intuition G regarding X. Moral consistency reasoning brings out that this goes along with, or is facilitated by, the *elimination or weakening* of the former intuition F regarding X.

Campbell and Kumar also provide one of the clearest statements yet of the idea that not only first-order judgments, but also *assessments of moral relevance* are based on intuitive, emotional responses.[769] First, consider an element of Campbell and Kumar's account that might have no counterpart in the intuitionist model, namely the negative affective response to conflict between intuitive evaluations of cases that do not appear relevantly different. This response is supposed to be an automatic "moral disapprobation [felt] toward individuals, including oneself, when they exhibit moral inconsistency."[770] It is not quite clear how Campbell and Kumar conceive of this mechanism. Is it a response specific to 'moral inconsistency', or an emanation of the affective mechanisms that produce first-order moral judgments and assessments of moral relevance? The second interpretation could invoke this passage: "[Negative affective] [r]esponses to apparent inconsistency, we hypothesize, are generated by the very same norms that produce our responses to the target and base situations."[771] However, another statement suggests a response specific to (moral?) inconsistency: "[I]f none of the norms in system 1 are activated—if the difference is not perceived as morally relevant, thereby engaging system 1—system 1 issues in a negative affective response."[772] Note that impressions of irrelevance here result from the *failure* of a given factor

769 For an earlier formulation of this idea, see Huppert (2010).
770 See Campbell & Kumar (2012), p. 290.
771 Ibid., p. 293. 'Norms' can be understood as regularities in the co-occurrence of specific situation-types and specific intuitive moral judgments. 'Target' and 'base' situations are the iudicanda under consideration.
772 Ibid., p. 291.

to activate 'norms', i.e., intuitions, anchored in system 1. I believe that this thought can be fruitfully combined with findings about the role of emotions in moral cognition in order to grasp the normative significance of moral psychology (see chapters 8.2 and 9.2).

As I have tried to show, different ways in which intuitive responses to specific situations and actions can arise and change are not equally shaped by system 1 or system 2 processing: Intuitive responses can be consciously (system 2) modified through input reconstruction (appraisal shifts) and input selection. Socialization can yield intuitive responses that are not innate, but also not the result of prior system 2 processing; implicit learning can form intuitive responses that might be more fully sensitive to the relevant consequences of our actions than conscious processing and application of moral rules can ever be. Sometimes, system 2 recognizes similarities between cases and leads to a revision of individual intuitive responses in order to resolve inconsistency with other intuitions. Presumably, researchers will discover even more mechanisms that affect moral evaluations. If, however, system 1 and system 2 indeed often *jointly* shape moral judgments, are positions that attribute special authority to *specific modes of processing* or to *processes supposedly unaffected by evolution* still viable?

# 6 Modules, Innateness, and Sources of Disagreement

This chapter further explores the evolutionary psychological perspective on morality and moral cognition. I examine how environmental influences affect the development of the mind. Assuming that evolved psychological mechanisms (EPMs) dealing with different problems produce outputs of discernible experiential qualities, I explore how the modular architecture of the brain might account for a particular effect caused by efforts to describe or explain moral-psychological phenomena: Scientific language is likely to be processed by psychological mechanisms *other* than those being investigated.

Research on social exchange cognition, for instance, suggests that moral judgment about different subject matters falls back on many separate specialized psychological mechanisms. This observation might explain why no *single* moral principle proposed so far produces judgments in line with our intuitions across many different contexts. It is also in line with the evolutionary-psychological idea of a modular mind consisting of many problem-specific evolved psychological mechanisms. In chapter 6.2, I discuss several arguments enlisted by Jesse Prinz against the idea that morality is in some sense innate: Firstly, an evolutionary account of morality could suggest that at least some moral principles exist in *all* cultures. While Prinz argues that *no* rule is universal, I claim that plausible suggestions for universal aspects of morality are still on the table, and that innateness of aspects of morality does not preclude extensive learning in the formation of full-fledged moral attitudes. Moreover, innate aspects of morality need not be concrete *principles* in order to have significant influence on the shape of morality. Prinz also argues that cognitive capacities involved in social exchange, such as a capacity to detect cheaters, do not in fact depend on a specialized, evolved psychological module. Rather, people who have deficits in social exchange reasoning have deficits also in solving other tasks. In my view, this is not a conclusive argument against innateness or modularity. Specific cognitive capacities could well be involved in several evolved modules. Another argument against innate components of morality brought forward by Prinz is that discontinuities in behavior between humans and nonhuman primates outweigh continuities, mainly because these primates have no *concept* of morality and thus cannot act from moral reasons. I respond that Prinz's understanding of morality is too demanding; it excludes many issues that are part of the moral domain. Using a broader notion of morality, similarities between human and nonhuman primate behavior become apparent and point to significant innate aspects of morality. Finally, rejecting evolutionary-

psychological accounts of morality on grounds of parsimony fails to take into account its explanatory power.

Further difficulties for attempts to formulate universal moral principles arise from differences between *individual* minds that result from variation in genotypes, in combination with idiosyncratic experiential input: No two brains are exactly alike; accordingly, they can differ in their moral evaluations. Both concretion and weighting of the 'moral foundations' vary across cultures and individuals. This great diversity not only explains why universal principles are unlikely to emerge, but also why disagreement occurs.

## 6.1 Social Exchange Cognition: Are There Specialized Evolved Mechanisms?

Chapter 4 was an attempt to illustrate the complex relations between emotions and morality, both in terms of how different emotions shape concerns at the basis of morality, and in terms of the motivational effects through which emotions regularly produce behavior that relates to moral codes in specific ways. What can we learn from these considerations regarding the determinants of moral judgment? Apparently, emotional responses to particular iudicanda affect what we consider morally relevant. There are good reasons to believe that the psychological mechanisms at the basis of these responses are evolved adaptations: The emotions considered in chapter 4 have been observed in societies all over the globe, and they appear to solve problems that regularly arise when creatures that advance their own inclusive fitness interact.[773] Cohabitation in highly social species requires that individuals have drives or motives that generate prosocial behavior, i.e. behavior that caters to the needs and desires of conspecifics. Without such motives, the ubiquity of conflict would render social existence disadvantageous.[774] Mutations that connect social-cognitive abilities with affective responses providing such motives spread due to their evolutionary advantageousness. It seems unlikely that humans navigate a type of existence that shares many structural challenges with the modus vivendi of nonhuman primates and other related social species *without* recourse to the specifically adapted, motive-generating mechanisms already available. Consequently, judgments of moral relevance rest on appraisals that elicit 'moral' emotions.

As research on social cognition indicates, at least some of these appraisals are highly specialized. Evolutionary psychologists have been studying problems of social existence

[773] See Moll et al. (2008b), p. 4, Zimbardo & Gerrig (2008), pp. 456–457, Evans (2001), pp. 4–17.
[774] A view developed in some detail in Kitcher (2011), chapters 1 and 2.

extensively, particularly social exchange. Although social exchange is omnipresent in contemporary societies, it is not a matter of course from an evolutionary point of view. Any change in behavior first appears in isolated instances. In order to become a species-wide adaptation, a new trait has to be more successful than extant alternative behaviors *even* if it is present in only a few individuals. It is not sufficient that a new strategy outcompete alternative behaviors in some hypothetical scenario; it has to be advantageous in the *actual* conditions it emerges into, including the behavior of conspecifics and other animals. How can cooperation, or exchange of goods or services in particular, stabilize in a noncooperative society *in spite of* the risk of exploitation faced by co-operators?

This question prompted evolutionary psychologists to look for so-called cheater detection mechanisms. John Tooby and Leda Cosmides conducted a series of experiments that built on the work of psychologist Peter Wason. Around 1966, in response to the empirical logicism of Piaget, who held that thinking develops to follow the rules of classical logic, Wason had devised an experimental setting to investigate whether human reasoning abilities correspond to ideal hypothesis testing (the *Wason selection task*).[775] More specifically, he examined reasoning about conditionals of the general form 'If P, then Q'. Typically, a setting contains four cards, each of which states either P or ~P on one side, and either Q or ~Q on the other. The cards face the subject with P, ~P, Q, and ~Q respectively, and the task is to turn around those and only those cards which have to be checked in order to find out whether the conditional rule is valid. Since the rule is violated whenever a card shows P on one side and ~Q on the other, the P- and the ~Q-cards have to be turned. Originally, experimenters chose rules like "If a card has a vowel on one side (P), then it has an even number on the other side (Q)" and cards displaying *a*, *b*, *2*, and *3* on their front. In this case, *a* (P) and *3* (~Q) have to be checked. Most subjects, however, decide to turn either only *a*, or *a* and *2*, even though the *2*-card (Q) cannot violate the rule: It is not stated that *only* those cards with a vowel can have an even number on the other side. Typically, only about 10 % of the subjects get it right.[776]

Why are these tasks so difficult, even though well-developed logical capabilities would have certainly been useful in all kinds of environments? Possibly, it is because the human brain did not evolve in response to the abstract *logical structure* of adaptive problems. If so, to what *does* it respond? Experimental research refuted several potential answers. In particular, results did not improve when rules involved *familiar* content and relations (such as a

[775] See Cosmides & Tooby (2008), p. 60 and Elqayam & St. Evans (2011), p. 234.
[776] See Buss (2008), pp. 273–275.

disease causing specific symptoms) instead of arbitrary relations between numbers and letters. Even when test persons were told *explicitly* that P&~Q violates the rule, only 25 % responded accurately.[777] A more promising explanation emerged in the 1980s, when significantly better results were reported from experiments in which the conditional rule expressed what a person is *obligated* or *entitled to* in a given context.[778] The so-called 'drinking age problem' is a well-known example: Test persons instructed to picture enforcing the following rule in a bar: "If a person is drinking alcohol (P), then he or she must be twenty-one years or older (Q)". There are four people in the bar: someone drinking beer (P), another person having soda (~P), a sixteen-year-old (~Q), and a twenty-five-year-old (Q). In contrast with the other settings, most subjects correctly check ~Q (the sixteen-year-old) and P (the person drinking beer). Again, the result allows for several explanations. Since most test persons were students, researchers surmised they might be familiar with violations of this particular rule. However, other tasks featuring unfamiliar conditions (e.g., "If a man eats cassava root, then he must have a tattoo on his face") elicited similarly high percentages of correct responses.[779]

*Social contract theory*, developed by Leda Cosmides and John Tooby, explains the content dependence of performance in Wason selection tasks by referring to modularity and very specific adaptive problems.[780] Tooby and Cosmides asked themselves which abilities successful participation in social exchange requires. In the context of their theory, a social contract is a conditional relation between a *benefit* and a *requirement*: If you take the benefit (P), then you fulfill the requirement (Q). They hypothesize that human beings need a *cheater detection mechanism* to avoid exploitation by someone who takes the benefit without fulfilling the requirement (or by someone who does not provide the benefit when the requirement was met). Maybe rule-violation detection is better in some types of Wason selection tasks because particular rules tap into psychological mechanisms that evolved to detect these cases of cheating. According to Tooby and Cosmides, results improve whenever subjects interpret P and Q as benefit and requirement respectively; they then intuitively check those who have taken the benefit (P), and those who do not fulfill the requirement (~Q); e.g.,

777 See Cosmides & Tooby (2008), p. 63.
778 See ibid., p. 64.
779 See Cosmides & Tooby (2002), p. 101.
780 See Buss (2008), pp. 270–276.

those drinking an alcoholic beverage and those under twenty-one.[781] These results, and further investigations that established significant performance differences between Wason tasks of the social-contract kind and the abstract kind across various cultures, indicate that the human mind is not by default equipped with capabilities of abstract logical reasoning.[782] Additional evidence suggests that the processes involved in solving social-contract-related tasks are highly specialized. Experiments showed that schizophrenic subjects (a condition associated with frontal lobe dysfunction) were able to reason just as well as nonschizophrenics on social-contract problems, but showed severe deficits when dealing with abstract logical problems.[783] Some authors hypothesized that good rule-violation detection is *not* due to the specific components of social contracts, but triggered by conditionals that, in a broad sense, state what somebody 'ought to do'. Such *deontic* conditionals include, for example, precautionary rules ("If you engage in a hazardous activity such as X, you must take proper precautions such as Y"). This suggestion is compatible with the observation that psychopaths reason worse than nonpsychopaths do on both precautionary and social-contract rules.[784] However, there are indications that not *all* deontic conditionals are processed by the same mechanisms: For instance, an individual suffering from damage to the orbitofrontal cortex[785] and the amygdala[786] was able to reason correctly on precautionary rules, but *not* on social-contract problems.[787] Moreover, subjects tend to differentiate between *accidental* and *intentional* violations of social contracts, but do not so differentiate with respect to violations of precautionary rules.[788] These results indicate that psychological mechanisms employed in conditional reasoning about social contracts do not correspond to abstract logic, but are instead adapted to specific problems and associated with specific brain areas.

Some scientists have challenged these arguments for an *innate* capacity to navigate social exchange. Jesse Prinz argues that, while obligations and other rules might appear similar,

781 Subjects taking the perspective of someone fulfilling the requirement seem to interpret the social contract as "If you fulfill the requirement (Q), then you take the benefit (P)". Accordingly, they check for violations by choosing (Q&~P). See ibid., p. 275. It has been argued that Cosmides and Tooby's task is subject to framing problems that make the correct answer easier to grasp in the social task. See Dunbar et al. (2007), p. 186.

782 See Buss (2008), p. 274.

783 See Cosmides & Tooby (2008), p. 85.

784 See Ermer & Kiehl (2010).

785 Located behind the lower part of the forehead, above the eyes, presumably involved in decision making.

786 Nuclei located within the medial temporal lobes, relevant for emotional reactions.

787 See Buss (2008), p. 275.

788 See Cosmides & Tooby (2008), pp. 102–104.

they are in fact very different, and that these differences can explain the differences in performance without reference to innate psychological mechanisms.[789] In particular, he contrasts social obligations with observations of correlations or causal relations, expressed in the form of conditionals (e.g., "If something contains alcohol (P), it will get you drunk (Q)"). In his view, we look for violations of such rules by checking cases in which P obtains (alcoholic content) and *not* by thinking about all beverages that do not make us drunk (~Q), because human beings learn about correlations and causal connections by generalizing data from *positive* cases. Social rules, in contrast, have to be enforced, and what is relevant for enforcement are not individuals who *conform* to the rule, but those who do *not* comply. When children learn social rules, they do so by experiencing the consequences of noncompliant behavior. Thus, according to Prinz, we learn social rules by attending to counterexamples, while knowledge about correlation and causation is acquired through observation of positive cases and generalization from these. Checking for ~Q with respect to correlational or causal rules, he claims, would be like learning that cookies are sweet by checking that non-sweet things are not cookies.[790] In his view, this tendency to learn about associations is a more likely innate capacity than a specialized capacity to think about social obligations.

While Prinz's explanation shows how different ways of reasoning about similar problems could result from differences in the environment (i.e., whether positive or negative cases are more instructive), I do not find his argument conclusive. In particular, it does not explain the performance difference sometimes observed between checking precautionary and social rules. Arguably, we learn precautionary rules by attention to cases of *non*compliance just as we learn social rules. If there were indeed a common learning mechanism behind the validation of both kinds of rule, no such difference would be expected. There might be subtler differences in how we learn to reason about social and precautionary rules that explain the differences in performance, for instance an attention to mental states (intentionality) with regard to violations of social rules, but not precautionary rules. In my view, the evidence quoted above certainly indicates that mental mechanisms *can* be highly specialized (modularity), and that specialization manifests in different neural substrates. A claim that some mental mechanism is innate can be supported by observations of that mechanism in related species and very young children, and is consistent with its presence in many or all cultures. The existence of stable exchange relations in many social species is certainly compatible with an innate preparedness to *learn* social exchange cognition. Other candidates for

789 See Prinz (2012), pp. 314–316.
790 See ibid., p. 316.

mechanisms evolved to avoid exploitation include the use of markers to differentiate trustworthy in-group members from out-group members, anger at and motivation to punish deviants, and the use of gossip to evaluate the trustworthiness of others.[791]

What is the upshot? Present knowledge about the workings of EPMs is certainly scarce in view of the amount of operations performed by the mind. However, the case for at least a certain degree of modularity seems convincing, and I believe that the findings mentioned offer themselves as starting points for a more general reflection on moral cognition. Precautionary rules and social contracts are two instances of a wide variety of contexts in which we use deontic vocabulary such as 'ought'. "You ought to do X", depending on context, might convey that X is required for attaining physiological goals (like satisfying hunger), or social goals as diverse as helping a friend, maintaining a good partnership, avoiding ostracism, manipulating others, etc. It can express moral obligation or prudential advice about how to reduce risk; generally, it can relate to all kinds of goals.[792] If psychological mechanisms are in fact often as specific as the findings regarding social exchange reasoning indicate, it suggests itself that the issues these various 'oughts' refer to are, at least in part, processed by discernible psychological mechanisms with different (though possibly overlapping) underlying neural substrates. More to the point, even the subset of topics we address as *moral* issues may in fact involve *distinct* problems, dealt with by disparate EPMs. Since natural selection shaped the human brain and thus affects how we think to a large degree (notwithstanding the importance of individual and cultural learning), it seems implausible that processes dealing with what a person 'ought' to do in matters ranging from, for instance, social exchange to sexual relations with relatives, should bear much resemblance. However, moral philosophy often treats conduct in exchange- and sexual relations, as well as in many other contexts, as subject to the same standard of evaluation, and many authors aspire to identify a *single* principle at the base of *all* moral obligations.

## 6.2 Scrutinizing Innateness Claims about Morality

According to Jesse Prinz, the view that "[b]asic moral values are moral sentiments directed towards various acts"[793] is neutral with respect to the question of innateness. However, he argues that four main arguments put forward by evolutionary psychologists in favor of innate morality are flawed. The following subsections address each of them separately.

791 See Dunbar et al. (2007), p. 184, Boehm (2012), p. 74.
792 See Cosmides & Tooby (2008), p. 59.
793 Prinz (2012), pp. 308–309.

### 6.2.1 Universal Moral Rules

If morality were innate, one would expect that at least *some* specific moral values or principles be in place in *all* human societies. The first thing to note about this conjecture is that innateness is *a* possible reason for universal presence, but not the only one. Human customs of producing clothing and building shelters are universal, but hardly innate. A more nuanced conjecture states that the *degree of variation* between moral practices in different societies should differ, depending on whether morality is mostly learned, or mostly inherited. Prinz, skeptical of innateness claims, argues that even what appear to be the most basic moral principles, such as 'harming is wrong', are *not* valid in all societies.[794] For instance, there are several reports about small-scale societies that do not appear to be squeamish about hurting or killing innocent people. Prinz mentions various examples to illustrate that aversion to violence tends to be strongest regarding, and sometimes limited to, the respective in-group. The delineation of this in-group varies across cultures. Ancient samurai sometimes tested new swords by "slicing a random peasant in half"; tribes in precolonial New Guinea incessantly waged war against each other. The Yanomami of the Amazon basin define their in-group by the village they come from, and regularly engage in violent behavior against members of their own tribe from other villages.[795] The Yanomami also kidnap their wives-to-be and see nothing wrong with beating them. Head-hunting cultures such as the Illongot of the Philippines "will take a head to relieve stress"[796]. Witnessing violence has been a widespread pastime through the ages: Consider the gladiators of Rome, public torture in medieval Europe, or the growing viewership of 'Ultimate Fighting' broadcasts.

Prinz's arguments against substantial innate moral universals do not convince Paul Bloom. To illustrate why, Bloom discusses Prinz's dismissal of an innate aversion to harming. Prinz quotes many examples in which people are *not* morally disturbed by instances of harming, or even think that it is morally praiseworthy. Bloom responds that innate aversions to harming are certainly more complex than 'harming is wrong'. Prinz's reaction is to claim that it is more plausible that the wide variation observable in the evaluation of harming is the result of deliberation about which kinds of norms can make a society work. Bloom in turn rejects this idea by stating that many norms about harming do not actually function to make societies work. Nevertheless, he concedes that the question is ultimately empirical and

[794] See ibid., pp. 309–314.
[795] See Prinz (2007b), p. 274.
[796] Prinz (2012), p. 310.

to be answered with the help of cultural psychology, primatology, and developmental psychology.[797] Bloom presents evidence of a capacity for social evaluation in babies and toddlers, comprising preference for those who behave prosocially and aversion towards those who act antisocially, as well as a preference for those who reward prosocial behavior and for those who punish antisocial behavior.[798] This last observation is particularly informative because a mere preference for prosocial behavior cannot explain it, and because it is incompatible with a general, dominant aversion to harming.[799]

In my view, Prinz's counterexamples affect only very simple versions of the innateness hypothesis. Sophisticated innate emotional tendencies that take group affiliation and other factors into account might be less vulnerable to his criticism.[800] What is innate might not be the communalities across cultures that appear most salient to us. For instance, the relational models proposed by Rai and Fiske are not obvious, but certainly worth taking seriously. In fact, if they can explain many patterns in moral evaluations *even though they are rather unobvious* candidates for universal features, that might support the suspicion that they are innate, rather than acquired through explicit instruction.[801] Even so, a (hypothetical) innate tendency to have negative emotional responses to physical violence against human beings does not imply that this inclination is equally strong in everyone, immune to cultural influence, or always the dominant force in shaping our behavior. Remember, for instance, that moral foundations theory explicitly allows that individual foundations vary in importance and concretion across cultures and individuals. Such differences might very well be susceptible to nonevolutionary explanation, while the basic capability to be sensitive to *these* fundamental concerns, rather than others, can plausibly be attributed to evolutionary processes (because they presuppose cognitive capacities that require a certain kind of brain). This thought may hint at an understanding of innateness that can defuse the apparent conflict between Prinz's position and the one advocated here. Prinz argues that "with innate domains, there isn't much need for instruction. Innate traits emerge on their own. In the moral domain, instruction is extensive."[802] However, the notion of innate *learning* modules employed by Haidt does not render instruction irrelevant; it just holds that some things are more easily learned

797 See Bloom (2012), pp. 72–75.

798 Interestingly, there is evidence that punishing defectors and rewarding cooperators are associated with increased activity in the same brain areas (striatum and medial prefrontal cortex). See TenHouten (2009), p. 149.

799 See Bloom (2012), pp. 76–79.

800 Such as "By nature, human beings are inclined to dislike harm done to members of their group that have not angered or disgusted them." In-group and out-group might be separated by various criteria (color of skin or hair, sex, etc.) without much cognitive effort.

801 There are, of course, also implicit forms of learning.

802 Prinz (2012), p. 322.

than others are. Traits with a significant genetic component of this sort do not develop inevitably, but require specific environmental input. Given this weaker notion of innateness, Prinz might agree that large parts of morality are innate, or shaped by evolutionary processes.

The next candidate innate principle Prinz discusses is a preference for equal division of resources.[803] He argues that we should distinguish between division of resources among kin and sharing with nonrelatives. Prinz claims that innate parental affection for offspring and similar mechanisms can explain sharing with relatives, but maintains that this is not a *moral* phenomenon. I disagree. Failure to provide for one's offspring frequently counts as a moral issue. What might be more difficult to explain in evolutionary terms is why we are morally moved by the failure of *unrelated* individuals to care for or share with their kin. Prinz also believes that the universality of sharing with nonkin is hard to explain in evolutionary terms. He argues that sharing with nonkin is a cultural development, as becomes evident in the cultural variation in what is considered a fair share. As an illustration, he quotes a study in which subjects assumed the role of a CEO who has to distribute bonuses, and can do so according to *merit* (business performance), *need*, or give *equally* to all. As it turns out, Chinese give to each equally, Indians are guided by need, and Americans distribute according to merit.[804] Again, I do not believe these examples are conclusive evidence against innate contributions to norms regarding resource sharing with nonkin. For instance, the innate attachment system that presumably motivates sharing with relatives could also be triggered, maybe to a lesser degree, by distress of unrelated infants (or, to an even lesser extent, mature individuals). Reasoning could to make their similarity to our own children more salient and thereby strengthen emotional activation. In my view, characterizing friendly responses to the distress of nonkin infants as innate is justifiable. As for the variance in fairness norms, I repeat that variation does *not* preclude significant influence of evolved psychological mechanisms. Culture certainly affects which *specific* division people prefer. Nevertheless, EPMs that spawn preferences for a nonarbitrary, rather than a universally fixed pattern of resource division, or norms that depend on which relational model is active, remain unaffected by Prinz's argument. I am not implying that Rai and Fiske's relational models are clearly innate. However, innateness remains a possibility that deserves further investigation.

Is reciprocity an innate moral rule? Prinz reports that psychopaths reciprocate only if it is in their interest.[805] Accordingly, he claims, universal reciprocity does not appear to be

[803] See ibid., p. 310.
[804] See ibid., p. 311.
[805] See ibid., p. 313.

innate, nor is the psychopath's reciprocity moral. However, it seems to me that absence of a trait in a small fraction of the population is not a good argument against the innateness of that trait. All innate properties are subject to mutation and environmental influences. What about the majority of human beings that presumably feels guilty for not reciprocating much more often than psychopaths do? Prinz argues that these guilt responses cannot have evolved because guilt-free members of the population would have exploited guilty-minded mutants. I am not convinced. After all, humans feel guilty mostly if their counterpart has *already* provided some kind of benefit. Are those who take the benefit and walk away better off? Not necessarily. The flipside of an evolved emotional sensitivity to reciprocity might have been an inclination to get angry with those who fail to reciprocate, which provides protection against exploitation by imposing costs on cheaters. Guilt might have evolved as protection against the angry responses of interaction partners. Finally, Prinz reports that cooperation rates in an iterated prisoner's dilemma vary widely across cultures, and even respond to the name of the scenario.[806] Such variability of a behavioral trait can appear to speak against its innateness. However, the remarks made above with respect to equal division of resources apply here as well: Name changes and cultural particularities might determine which innate relational model is active, and thus what *level* of reciprocity is considered adequate. Prinz anticipates such objections: He holds only that no *specific* moral rules are innate, since the probability of universality of a given rule increases the more abstractly the rule is. There might be a general rule against harming *some, but not all* 'innocent' people. One might object that rules that, for instance, incorporate in-group/out-group distinctions to determine the permissibility of harming, are *more,* rather than less, specific. Even so: Significant evolutionary influence on moral rules does not preclude that culture *also* shapes old as well as new moral concerns. Commonalities of the inputs and outputs processed by evolved psychological mechanisms might not always be obvious; nevertheless, these mechanisms co-determine the shape of morality. Indeed, anthropologists have listed behaviors that appear to be condemned universally (subject to cultural variation with respect to severity of punishment), such as "undue use of authority, cheating that harms group cooperation, major lying, theft, and socially disruptive sexual behavior."[807]

[806] See ibid.
[807] Boehm (2012), p. 34, see also ibid., p. 46.

### 6.2.2 Cheater Detection

Evolutionary psychologists Tooby and Cosmides believe that the evidence for a cheater-detection module supports the idea of innate aspects of morality. Prinz attempts to undermine these claims by providing alternative explanations that involve learning mechanisms for those experimental and observational results (see chapter 6.1). Like universality, selective deficits in moral cognition could be evidence for innate morality. If morality-specific modules exist, there should sometimes be individuals lacking just the capacities these modules provide, but not others. The individual, mentioned in chapter 6.1, who showed deficits in reasoning about social obligations, but *not* in reasoning about precautionary rules, is such a case. Prinz argues that this individual's impairments were probably not limited to morality. His performance in most other areas was simply not tested. The specific lesions this patient suffered are associated with deficits in emotional functioning more generally, and since morality very much depends on emotion, it is not surprising that morality was affected. Similarly, psychopaths, contrary to the claims of James Blair, have deficits beyond morality, such as difficulties in appreciating art and music, and decision making.[808] What exactly do these arguments establish? They show that individual structures in the brain are not involved *exclusively* in psychological operations that we consider part of morality. However, proponents of innate moral mechanisms need not claim that they are. EPMs could consist in specific patterns of distributed neural activity that recruit several localized structures. If one of these structures is involved in several psychological processes, a lesion to that region can result in multiple deficits. Moral traits or behaviors could nevertheless emerge from evolved neural structures. On the other hand, even if some psychological process appears to be located in a specific brain region or specific distributed networks, that does not imply that it is innate rather than learned. It might just be the case that a particular kind of learned information that is relevant for moral functioning is stored in that area or network. Do these considerations show that the nature-nurture debate is pointless, since all learning must be reflected physically in the brain? That depends on the range of things these neural structures can learn. Garcia and Koelling's experiments show that the set of phenomena rats easily link to nausea is limited (see section 1.2.2). What is innate is not the association between particular foodstuffs and nausea, but rather the *ability to learn* that kind of association with respect to foodstuffs, and not flashes and buzzers. Arguing for innate features of hu-

[808] See Prinz (2012), p. 317.

man morality does not require positing innate moral *principles*. Innate, probabilistic boundaries on the range of what can count as morally relevant are more plausible, especially given that the human brain is much more plastic than that of other animals.[809]

### 6.2.3 Moral Apes?

Further arguments in favor of innate morality point to interesting behavior in animals, particularly nonhuman primates.[810] Vampire bats (which are not primates) display altruistic behavior, mainly towards relatives. Chimpanzees share food and console upset cohabitants, among other prosocial behaviors. Precursors of moral behavior in animals that lack language would be strong evidence for innate components of morality in humans. However, Prinz considers the continuities between animal behavior and human morality insignificant compared to the discontinuities. More importantly, he argues, even animal behavior that *appears* to share ancestry with human morality is not truly *moral*, since it is not done *for the right reasons*.[811]

Are behavioral continuities between humans and their closest relatives insignificant when it comes to morality? First, consider which behaviors in nonhuman animals could count as 'building blocks' of morality, and which traits seem to be uniquely human. Sharing of food and other resources, helping behavior, and reciprocal exchange occur in nonhuman primates and other species.[812] Chimpanzees and other social animals repair relationships that have been damaged by aggression by what Frans de Waal calls 'reconciliation', a friendly contact between parties that have been in conflict shortly before.[813] De Waal also holds that chimps and bonobos, and maybe dogs, have experiences very similar to guilt. If this claim about chimpanzees and bonobos were correct, it would imply that this capacity is at least 5 to 7 million years old, for this is when the last common ancestor of humans, chimpanzees, and bonobos (sometimes referred to as *ancestral pan*) supposedly lived.[814] I have argued in chapter 4 that many emotions whose elicitors and action tendencies shape the moral domain are *not* psychological innovations that occur only in humans, but stem from EPMs that

809 See Hüther (2011), pp. 53–61.
810 See Prinz (2012), pp. 318–320.
811 See ibid., p. 319.
812 See Prinz (2007b), p. 273.
813 See De Waal (2005), pp. 19–20.
814 See Boehm (2012), pp. 90–91.

solve similar problems in animals, combined with the cognitive capacities typical of humans.[815] This seems true of anger, contempt, some forms of morally relevant disgust, shame and embarrassment, guilt, pride, gratitude, and affective empathy.

An impressive variety of morality-related behaviors has been observed in nonhuman primates, and partly also in other social mammals: Researchers mention retributive behavior, reciprocal helping, consolation, and conflict mediation.[816] There is also evidence of problem solving, mother-child bonding, special treatment of the injured or disabled, awareness of group membership, attempts at deception and anger at their discovery.[817] It has been claimed that, when presented with an opportunity to benefit a friendly group member at no cost to themselves, chimpanzees do not grasp it.[818] Frans de Waal and colleagues, however, changed the experimental procedure and found instead that chimpanzees *do* act in the interest of others on such occasions.[819] Further observations include emotional contagion, violence inhibition, incest avoidance, awareness of one's reputation, aversion to inequity (disadvantaged parties become agitated, but even the advantaged party is more likely to reject its higher-value reward)[820], and third-party intervention.[821] Advocates of relational models theory (chapter 3.2.3) suggest that chimpanzees structure relationships according to community-sharing (CS) and authority-ranking (AR) models. There is some debate regarding whether they employ the equality-matching (EM) model, which requires equal distribution of some good, and market pricing (MP), which implies a sense of proportionality.[822]

On the other hand, important features of morality probably are uniquely human. In terms of behavior, this includes regular third-party punishment (i.e., an observer who is not affected by a transgression and not closely related to the victim imposes costs on the transgressor at a cost to himself), widespread reciprocal interaction with nonkin and helping

[815] Damasio (2005), p. 47 believes that "there was a biological blueprint for the intelligent construction of human values, and that the biological blueprint was present in nonhuman species and early humans. We also believe that a variety of natural modes of biological response, which include those known as emotions, already embody such values. They too were present in nonhuman species and early humans."

[816] See Cela-Conde (2005), p. 11, De Waal (2013), p. 128. Even ravens respond to distress in conspecifics. See ibid., p. 6.

[817] See Verplaetse et al. (2009), p. 22, Greene (2002), p. 155. See Brosnan (2012) for an overview of research on responses to inequality in nonhuman primates.

[818] Capuchin monkeys, in contrast, "take their partner's outcomes in to [*sic*] account". Brosnan (2014), p. 93.

[819] See De Waal (2013), pp. 118–121.

[820] See ibid., pp. 233–234. De Waal (2014), pp. 196–197 reports that "advantageous" inequity aversion on the part of the agent who profits has been observed in some apes, but not in monkeys.

[821] See Verplaetse et al. (2009), p. 31, Brosnan (2014), p. 89. Suppression of inbreeding is present in species as different as fruit flies, rodents, or primates. See De Waal (2013), p. 71.

[822] See Sunar (2009), p. 454.

behavior towards nonkin without a prospect of reciprocation, as well as socio-moral disgust.[823] While gorillas sometimes, and chimpanzees and bonobos regularly, form coalitions to punish or suppress bullying by alpha types, human foragers engage in counterdomination to such an extent as to render their groups egalitarian.[824] The vast majority of prosocial actions in chimps is directed at in-group members and occurs in dyadic interactions. Humans are frequently concerned about third parties, take their moral rules to be valid even in faraway places, and are less likely to miss opportunities to help friends at no cost to themselves.[825] Reputation building is also much more widespread in humans.[826] Chimpanzees depend less on each other than human foragers do because they do not hunt large game; they engage in collective action less frequently, and cooperate less intensely.[827] Such differences presumably result from the superior cognitive capacities of humans, such as the ability to represent future states that differ significantly from the present and thereby explore the potential consequences of one's actions, conceptual abstraction, an understanding of how others think, and the ability to learn numerous attitudes and behaviors through imitation.[828] These new or highly advanced abilities seem nevertheless to operate on the motivational machinery that has developed throughout the evolution of the human species and is present at least in similar form in the great apes.[829] While behaviors mentioned in accounts of animal protomorality have more of an intuitive character, humans can think about these responses, debate them with their conspecifics, and alter the corresponding norms.[830]

Back to Prinz's arguments: He claims that the dissimilarities between animal and human behavior with respect to morality clearly outweigh the similarities, and that animals do not act morally because they have no *concept* of morality and therefore cannot act from moral motives.[831] Both claims depend on Prinz's particular notion of morality. If only behavior done based on the conviction that it is what you *morally ought* to do (a motivation-based notion) counts as moral behavior, then it is plausible to say that *no* animal acts morally: Only

[823] See Moll et al. (2008a), p. 162, Haidt (2001), p. 826, Verplaetse et al. (2009), p. 32.
[824] See Boehm (2012), p. 96.
[825] See Prinz (2012), p. 319.
[826] See De Waal (2014), pp. 199–200.
[827] See Van Schaik et al. (2014), p. 81.
[828] See Moll et al. (2008a), p. 175, Verplaetse et al. (2009), p. 32, Prinz (2012), p. 321. Many of these abilities seem to involve the neocortex, that is, the evolutionarily most recent parts of the human brain.
[829] A thought present already in Darwin's famous quotation "[A]ny animal whatever, endowed with well-marked social instincts, the parental and filial affections being here included, would inevitably acquire a moral sense or conscience, as soon as its intellectual powers had become as well, or nearly as well developed, as in man." Darwin (1871), p. 98. Boehm (2012), p. 7 believes that this view degrades conscience to a mere by-product; I doubt that this is the only conceivable interpretation.
[830] See Damasio (2005), p. 51.
[831] See Prinz (2007b), p. 262.

humans have the cognitive capacity to entertain the notion of a 'moral code' and evaluate whether their behavior conforms to that code. Prinz is even more specific: He claims that universal validity of norms is an important part of human morality, and that guilt is exclusively human since it can only arise based on the impression that one has violated a *moral* rule.[832] Since chimpanzees, according to Prinz, lack a capacity for guilt (a claim doubted by de Waal, as mentioned in chapter 4.5.1.2)[833] and do not care about the universal validity of rules, what appears similar to human moral behavior is in fact motivated by fear, or by a simple dislike of seeing others suffer.

Let us concede that even if nonhuman primates are capable of an emotion very similar to guilt, its triggers probably do not involve a cognitively demanding concept of morality. It is, however, less certain that the motivational mechanism is qualitatively different from what occurs in humans. Guilt has negative valence; it is unpleasant. Thus, efforts to make amends or to conform to the relevant rules in the first place might well be strategies to avoid this experience. Yet such 'egoistic' avoidance motivation seems to be what *disqualifies* prosocial acts done out of fear from counting as moral on Prinz's account. The debate about the existence of psychological altruism (chapter 4.4.3) shows that it is quite challenging to determine whether specific actions are done to satisfy altruistic ultimate desires or to 'egoistically' avoid unpleasant experiences of moral insufficiency. As Philip Kitcher points out, demanding that genuinely *moral* motivation be free from influences of fear and prudential calculation would imply that "most of the people who have ever lived" were no moral agents, since their ethical practices were based in religious beliefs that motivated compliance through fear and awe.[834] Prinz presupposes much more idealized motives behind human 'moral' action than explanation of human behavior requires there to be, and than would be plausible to assume for much behavior typically considered morally praiseworthy. Since I am concerned not with an idealized notion of morality whose practical relevance is uncertain, but with all motives that shape social behavior and affect our moral evaluations, I am concerned with a broader spectrum of psychological mechanisms than Prinz.

---

832 It is curious that Prinz would require *all* moral rules to be universally valid. Sometimes, he relies on Shweder's three ethics to explain that morality is not limited to the ethic of autonomy. However, the ethic of community arguably revolves around hierarchy and roles. One could argue that hierarchy and roles are established *precisely* by norms that are *not* universally valid, at least in a certain sense of universality.

833 However, de Waal would probably agree that chimpanzees do not experience guilt because they realize that they failed to do what they *morally* ought to do.

834 See Kitcher (2011), pp. 80–81.

A similar point can be made regarding the capacity to understand moral rules as generally valid. As the discussion of the moral/conventional framework and Haidt's cultural-psychological research has shown (chapter 3), general validity might not be as typical of moral judgments as frequently assumed in philosophy and presupposed in the moral/conventional tradition (remember the example regarding the punishment of sailors today and hundreds of years ago). Again, it is doubtful whether most people's understanding and practice of morality conform to Prinz's demanding criteria. Given my more liberal understanding of what constitutes morality, the differences between humans and their most advanced relatives appear less stark. Yet, even if neither the absence of fear or avoidance motivation nor the universal validity of rules are good criteria to distinguish parts of human morality from certain motivations in nonhuman primates, the mediating role of the notion of a moral code remains an important difference. However, does it justify the claim that the evolutionarily shaped aspects of morality are relatively insignificant? Prinz argues that

> [w]e share a non-moral tendency to share, to help, and to reciprocate with our primate cousins, and these tendencies become objects of moral praise in us. We *learn* to view these things as good. We can develop moral attitudes toward other behaviors, but the behaviors emphasized by evolutionary ethicists are typically central to human life.[835]

As discussed with respect to the role of emotions in chapter 4.3.5, I agree that there is a difference between having a concept of moral rules and acting prosocially without prior moral reflection. I do not want to claim that animals can act from moral motives in the same way humans sometimes do. However, with a look to a descriptively adequate account of the actual motivations involved in human behavior within the domain of morality, it seems difficult to establish a standard by which to judge the *discontinuities* between humans and their evolutionary relatives 'more significant' in shaping contemporary systems of morality. The presentation of these issues depends on whether your theoretical predilections are nurturist or naturist, and which aspects of morality you consider particularly important. Proponents of innate aspects of morality can argue that even sophisticated notions of moral rules as quite universally valid, which require cognitive capacities that nonhuman species lack, still depend crucially on attachment systems, empathic capacity, perception of hierarchies, and associated (proto)emotional mental states that can plausibly be attributed to chimpanzees and other species. A mind with the emotional setup of a chimp and cognitive capacities comparable to ours, in a setting in which it is aware of and potentially interacting

[835] Prinz (2007b), p. 273.

with large numbers of individuals of the same species outside her own group, might well develop concerns for third-party interaction and other particularities Prinz conceives of as qualitatively different from nonhuman primate behavior.

On the other hand, it is quite obscure what beings with cognitive capacities comparable to those of humans would consider 'moral' if their minds lacked the (proto)emotional mechanisms humans apparently share with their close evolutionary relatives (e.g., anger, protoshame, (core) disgust, affection, etc.).[836] In short, the potential *content* of morality is circumscribed in large part by psychological mechanisms that we inherited from our evolutionary ancestors. Particularly with regard to the concerns about morality and moral judgment raised by evolutionary debunking arguments, I want to emphasize that even if no moral *principles* of the kind discussed in moral philosophy are innate, there are good reasons to assume that *which principles endure* depends on evolved emotional capacities. Even if, as Prinz claims, actions are moral only if they reflect the special normative authority of universal moral rules, and true guilt arises only from the impression that one has failed to act *morally*, the influence of evolved fundamental concerns on the *content* of these rules is still substantial. While Prinz's arguments against specific evolutionary-psychological claims seem insufficient to discredit the research program, they do clarify that universality and early emergence of traits are in principle compatible with *both* evolutionary-psychological and learning-based explanations. Similar behavior in humans and animals that lack comparable capacities for ontogenetic learning, in contrast, is a good indicator of an evolutionary origin.

### *6.2.4 Does Parsimony Favor Nurturism?*

In addition to the discussion of specific evolutionary-psychological claims, Prinz makes a methodological argument for parsimony and against assuming innateness for universal moral rules: Even if biological evolution might explain them, cultural evolution does, too, since most of the norms discussed are required for stable societies.[837] Prinz believes that, all else being equal, assuming fewer entities is a virtue in theorizing. Cultural explanations require only very general innate learning mechanisms, and are thus more parsimonious than evolutionary-psychological theories that postulate larger numbers of domain-specific modules. Is this a good argument? Jonathan Haidt has argued that widespread overvaluation of

[836] This thought-experiment presupposes that what is conceptualized as 'rational' or 'cognitive' (in the narrow sense) information processing can be separated from the emotional mechanisms. This might not be the case. The existence of computers fuels suspicions that it is possible; however, human cognitive architecture might work very differently even if computers can reproduce specific input-output couplings.

[837] See Prinz (2012), p. 314.

parsimony has spawned theories that are incapable of capturing the complexity of the phenomena they aim to explain. The attempt to reduce all of morality to concerns about harm and fairness is a case in point. Haidt does not oppose Ockham's principle: He assumes that moral foundations theory is *superior*, rather than equal, to the moral/conventional framework in terms of explanatory power. A more parsimonious theory might still be the better choice if theories are *equal* in explanatory and predictive power. However, worrying too much about parsimony can cause blindness to phenomena that do not fit a simple theory.

With regard to the examples Prinz discusses, it is, at this stage, very difficult to compare the explanatory and predictive power of the nativist model of moral psychology (evolved modular minds) with the nurturist alternative (general learning and cultural evolution). The extent to which evolutionary influences appear to shape morality also depends on your notion of morality. If your notion of morality emphasizes cognitive processing of general rules, a sense of duty that is separate from inclinations, or a similarly demanding criterion, then the evolutionary history of human brains can certainly seem less relevant. On a more inclusive view of morality, a view that captures a large variety of motivations that make people come up with and conform to rules that regulate social life, the continuities between related species and humans become obvious. To me, the inclusive perspective is more attractive because the alternative leaves out many of the psychological processes that are highly relevant for sociality, and ignores the phylogenetic background of these capacities. On such an idealized view, moral action may be quite a rare phenomenon. That is not what I am interested in. Most people, at all times and places, evaluate the behavior of others as well as their own, and these evaluations affect how they relate to each other. That is what I want to understand, and the narrow, idealized notion of morality will just not do.

## 6.3 Sources of Moral Disagreement: Genes, Culture, and Individual Experience

Modularity of the mind is not the only reason why moral cognition could be a more heterogeneous phenomenon than some philosophers seem to believe. There are also important differences between *individuals*. Inheritance, i.e., aspects coded for by genes, presumably plays an important role in the formation of human psychology. The concepts and structures in which we cognize and think are, however, also strongly influenced by environmental input. Evolutionary psychology does not claim that moral, or any other kind of thinking, is genetically determined. Rather, heredity and environment interact in manifold ways to form individual minds.

Any genetic predisposition depends on the environment for its expression. To take a simple example: Even though we have genes for muscles in arms and legs, their actual development depends on whether the environment provides sufficient nutrition. Genotypes[838] do not spread because they guarantee increased inclusive fitness *independently* of any 'outside' or nongenetic factors, but because they interlock with regularly occurring features of the environment in ways which, on average, lead to more successful reproduction. Hence, various patterns of interaction between genotype and environment in the development of the phenotype[839] are conceivable. For instance, psychological mechanisms can be context sensitive: Children learn whatever human language they are exposed to for a sufficiently long period during specific phases of their upbringing. Social contract EPMs appear to work with rules from very different cultures, as long as requirement and benefit are identifiable. Other EPMs might make us seek out input necessary for the development or calibration of other mechanisms.[840]

*Phenotypic plasticity* enables advantageous gene-environment interaction:[841] Some genes are selected for because they cause the development of different phenotypes corresponding to the specific requirements of the environment. For instance, polar bears from the same genetic stock develop fur of varying thickness, depending on the temperatures prevalent in the environment they live in.[842] Similar responsiveness to environmental conditions is presumably possible with respect to the mind. In general, many aspects of an individual's morality are likely heritable:

> On just about everything ever measured, from liking for jazz and spicy food to religiosity and political attitudes, monozygotic twins are more similar than are dizygotic twins, and monozygotic twins reared apart are usually almost as similar as those reared together […]. Personality traits related to the five foundations, such as disgust sensitivity […] or social dominance orientation (which measures liking for hierarchy versus equality; […]), are unlikely to be magically free of heritability. The "Big Five" trait that is most closely related to politics—openness to experience, on which liberals are high—is also the most highly heritable of the five traits […]. Almost all personality traits show a frequency distribution that approximates a bell curve, and some people are simply born with brains that are prone to experience stronger intuitions from individual moral modules […].[843]

838 The genetic makeup of an individual.
839 The set of observable characteristics of an individual.
840 See Tooby & Cosmides (2001), p. 15.
841 See Sober (1998), p. 132.
842 See Sober (1997), p. 543.
843 Haidt & Bjorklund (2008), p. 210.

Environmental influences such as explicit instruction, practice, behavior of adults and peers, or the media affect how these inherited preparations grow into sensitivities and virtues related to the six foundations. These processes can broaden or narrow the modules; moreover, they can strengthen or weaken the motivational force of individual foundations. Haidt et al. also mention the possibility that certain individuals are much more sensitive to potentially moral aspects of events and actions, and may perceive as morally relevant what others ignore.[844] Furthermore, evolved psychological mechanisms are affected by information stored in individual brains, so that every human mind is constantly influenced by a pattern of environmental input that is unique not only because individual experiences (input received) are unique, but also because they are processed by mechanisms whose workings are shaped by individual life history. Personal characteristics such as working memory capacity, sensitivity to reward and punishment, need for cognition, or personality traits such as extraversion cause differences in judgment.[845] Differences at both the individual and the cultural level occur also in controlled cognition, mental-state reasoning, and emotional responding.[846] Haidt et al. list several additional processes that could explain differences in moral judgment:[847] Some individuals gather more information than others do before passing judgment. Higher intelligence or increased need for cognition can facilitate post-hoc rationalizations (link 2 in the social-intuitionist model, Figure 4) and persuading others in reasoned arguments (link 3); they might also strengthen links 5 and 6. Those who are less responsive to reward and punishment, and think longer, could be less susceptible to the mere presence of certain judgments in others (link 4: social persuasion).

The moral foundations framework and relational models theory illustrate how and why emotional responses affect morality, and why a surprising variety of characteristics of iudicanda can appear morally relevant to different individuals. Every feature of a iudicandum that can be *framed* so as to be processed by mechanisms more or less closely connected to those that establish fundamental moral concerns has the potential to appear morally relevant. This *concretion* of fundamental moral concerns is, however, not the only source of variation in notions of moral relevance. Cultures, subcultures, or individuals also assign different *weights* to individual moral concerns. Haidt and colleagues investigated conservatives and liberals in the USA and found that liberals are mainly concerned with issues of care/harm, liberty/oppression, and fairness/cheating (where fairness as proportionality can

844 See ibid.
845 See Waldmann et al. (2012), p. 285.
846 See Young & Saxe (2011), p. 323.
847 See Haidt & Bjorklund (2008), pp. 210–211.

be traded in for liberty and reduction of harm). In contrast, conservatives attribute approximately equal importance to all six moral foundations (care concerns can be traded in for other, more important goals).[848] Such differences can be learned, but they can also stem from other sources: Since psychological mechanisms correspond to brain structures, all factors that affect the brain can affect these mechanisms, including environmental influences other than learning (e.g., nutrition, pollution), mutation, genetic recombination, etc. Such influences could result, for instance, in a particularly pronounced ability to feel compassion, or a greater susceptibility to purity concerns anchored in disgust. Singer's and fellow utilitarians' focus on welfare, fulfillment of personal preferences, and similar values mark a moral-psychological make-up dominated by concerns for care/harm, liberty/oppression, and fairness/cheating.

Kanai et al. found systematic differences between the brain structures of young adults that corresponded to their self-reported political orientation. While the right amygdala of conservatives was larger in volume, liberals had larger gray matter volume in the anterior cingulate cortex (ACC).[849] Correlational studies like this one do not establish causality: Do people adopt a political attitude because they have a specific kind of brain, or does their political attitude shape their brains? Kanai et al. suggest that larger ACC volume in liberal individuals might point to a higher tolerance for uncertainty, which facilitates the adoption of liberal attitudes. The amygdala is associated with the processing of fear, so larger volume in this area could imply an increased sensitivity to fear that conservative attitudes express.[850] However, functional interpretation of neural differences currently remains somewhat speculative. Brain regions are generally associated with *several* functions because the resolution of available brain imaging techniques is too low to differentiate between them. At this stage, there is no way of knowing whether the potential connection between a larger right amygdala and conservative attitudes consists in the processing of fear, or rather some other function which involves the right amygdala.[851] Even so, interpretation of neuroscientific findings is not arbitrary, but constrained by a requirement of compatibility and coherence with existing research.

---

848 See Haidt & Graham (2007), and more recently, Haidt (2012), p. 184. According to Lewis & Bates (2011), the expression of 'character(istic) adaptations' produced by personality traits and external factors depends on the interplay of adaptation and context (p. 548). They refer to the 'two-factor model of morality' by Graham et al. (2009).

849 See Kanai et al. (2011), p. 677. The anterior cingulate cortex is a part of the cerebral cortex surrounding the frontal part of the corpus callosum and associated with autonomic functions as well as error- and conflict detection.

850 See ibid., p. 678.

851 See ibid.

In sum, with the exception of monozygotic multiples, each individual's brain and mind are unlike any other for at least two kinds of reasons: Individuals have a unique genotype. Moreover, no two individuals experience the *exact* same environmental influences. Thus, just as individuals have different fingerprints, scars, or teeth, they have different brains and different minds. This is not to overemphasize variation among conspecifics. Certainly, the members of any species are similar in many respects. To a certain degree, emphasis on similarities and differences depends on discipline: Cultural psychology concentrates on the *differences* in emotions, thoughts, etc. between cultures and emphasizes context dependence, while evolutionary psychology seeks to explore the 'psychic unity of humankind'.[852] There might be incentives in behavioral science to focus on universality claims rather than explorations of diversity, since they offer simplification, are more elegant, promise more predictive power, and are easier to teach.[853] In the context of this thesis, however, focusing on diversity helps to understand, for example, why people can differ in their moral judgment *even if* they seem to agree on all 'objective' features of the iudicandum. It also clarifies why almost any attempt to capture the determinants of moral judgment in a phrase, an article, or an anthology is bound to meet with opposition from someone. Traditional accounts of these determinants can identify important aspects, but they are too simplistic in light of the innumerable parameters (including, of course, knowledge of ethical codes or theories) which influence moral verdicts, and underestimate the differences between individuals.[854] From this point of view, a complete explanation of an individual's moral convictions would include an incredible amount of information, and it would apply in its entirety only to this specific individual at this specific point in time.

Theories like the moral foundation framework and relational models enlighten the origins of moral relevance and, to some extent, the determinants of moral judgment. They also provide tools for understanding different cases of *ir*relevance-impressions. In some cases, particularly when determinants of moral judgments seem irrelevant to some, but not others, disagreement could be the result of differences in the concretion of moral foundations, or their relative weighting. *Un*controversial moral irrelevance, in contrast, is a sign for the psychological difficulty of relating the factor in question to fundamental moral concerns.

852 See Turiel (2006a), pp. 794–795. He somewhat ambiguously states that *both* "supposedly new disciplines" claim that the respective *opposite* paradigm was dominant in psychology, and is now being replaced.
853 See Rochat (2010), p. 107.
854 See also Wilson (1975), p. 564.

# Part III

—

# Philosophical Repercussions of Moral Psychology

# 7 How Normative and Metaethical Significance Depend on Psychological Facts

After this long discussion of moral-psychological research, let me recapitulate the different positions regarding the normative and metaethical significance of moral psychology presented in chapter 2. Peter Singer claims that empirical findings, in combination with evolutionary explanations, show that certain moral intuitions are not trustworthy, because they respond to differences between iudicanda that have no 'moral salience'. This view fits his general distrust in intuitions. Singer argues that moral judgment has to be more 'rational'. Moral psychology might provide a better understanding of what 'rational' means. Selim Berker disputes any direct normative significance of neuroscientific research. According to him, Joshua Greene's argument against deontological, emotional intuitions hinges on *normative* intuitions about the moral relevance of factors that affect judgment, and the respective experimental results do not speak to the adequacy of these intuitions. In Berker's view, neuroscience can at best show that brain areas whose activation is associated with errors in *other* domains of judgment are also active in *moral* judgment, thus prompting vigilance for similar errors in the moral domain. In response, Greene explicated his notion of the normative significance of moral psychology in more detail: Moral psychology provides information about factors that influence judgment. We can intuitively assess the moral relevance of these factors. If we deem them irrelevant, arguments from morally irrelevant factors, in his view, undermine the trustworthiness of the judgments in question. Moreover, moral psychology can help determine whether moral judgments are shaped mainly by emotional processes, have an evolutionary background, or are heuristic processes. All of these characteristics supposedly render judgments unreliable at least in fundamentally new situations. In such situations, we should rely on cognitive processing, and moral psychology enables us to be more precise about what that requires.

Guy Kahane worries that evolutionary debunking arguments might have global reach, if evolution indeed affects *all* evaluative judgments. In that case, evolutionary debunking would lead to global evaluative skepticism. In addition, he claims that evolutionary debunking presupposes moral objectivism (the view that the truth of moral statements is mind-independent), since it would otherwise be impossible to criticize *mental* processes that generate moral judgments. Sharon Street argued that evolutionary processes do indeed shape *all* moral judgments more or less directly. In her view, this makes moral objectivism implausible, since it either implies skepticism in case there is no relation between evolutionary

processes and moral truth, or a tracking account containing such a relation, which would be scientifically inferior to an 'adaptive-link' account of the connection between evolutionary processes and moral judgment. Victor Kumar and Richmond Campbell believe that moral psychology, in combination with intuitions about moral relevance, can identify uncontroversially irrelevant, but psychologically efficacious differences between iudicanda. In these cases, the respective pairs of judgments are *jointly* unwarranted. Further normative argument is required to determine *which* judgment should be discarded. Empirical research can provide neither relevance judgment nor normative argument.

Assessing these different positions regarding the normative and metaethical significance of moral psychology requires a descriptive account of judgments of both first (moral judgments about concrete cases) and second order (judgments about moral relevance). Authors often, implicitly or explicitly, point to the moral irrelevance of factors whose influence has been identified, be it at the proximal, psychological level of explanation, or on levels that are more distal. We therefore need to determine how reliable judgments of moral relevance are; this, in turn, seems to call for an understanding of how they come about. While it became clear in the discussion of Greene's position that impressions of moral relevance do much work in arguments for the normative significance of moral psychology, it was not always discernible at which level of explanation accusations of irrelevance aim. A more fine-grained picture of moral judgment could help explicate and understand these arguments. Identifying the factors that influence first-order judgment requires a close look at the mechanisms involved, including their development. As I will argue, the pertinent findings also tell us something about the mechanisms and reliability of judgments about moral relevance.

Moral-psychological research bears on all the philosophical positions regarding its normative and metaethical significance that I have presented. The scope of evolutionary influences is particularly relevant. Regarding the arguments of both Singer and Greene, it is crucial to know the extent to which evolution shapes first-order moral judgments, but also, as I will argue, judgments about moral relevance. If evolutionary influence is pervasive, then, as Kahane suggested, evolutionary debunking arguments could lead to global evaluative skepticism. The applicability of evolutionary explanations also affects Street's argument against objectivism. If both global skepticism and tracking accounts turn out to be implausible, moral-psychological findings might indeed have metaethical consequences. Prompted by Berker's analysis of Greene's position, I want to figure out whether typically consequentialist judgments, rather than characteristically deontological ones, are similarly susceptible

to evolutionary explanation. Whether evolution affects judgments about the moral relevance of factors is also crucial. Generally speaking, the descriptive account of moral judgment should provide a better understanding of relevance appraisals, based on which I can assess their reliability and the cogency of arguments that incorporate them. The reliability of these judgments affects Kumar and Campbell's position as well. Even though, in their view, moral psychology can show only that specific judgments are *jointly* unwarranted, but not *which* judgment is mistaken, the normative significance of moral psychology, manifested in the detection of joint unwarrantableness, nevertheless depends on judgments about moral relevance.

The relative importance of rational and emotional/intuitive processes is another main issue. Singer and Greene claim that 'rational' or 'cognitive' mental processes generate more adequate moral judgments than emotional or intuitive processes (in Greene's case, this is true at least with respect to fundamentally new iudicanda). Since this attribution of trustworthiness presupposes that rational/cognitive processes are *not* subject to judgment-shaping evolutionary influences, it is crucial to check whether this assumption holds. Greene is particularly suspicious of judgments dominated by emotional processes. Therefore, in preparation for an assessment of his arguments, I discussed the extent to which emotions shape first- and second-order judgment. Moreover, I investigated the degree to which the human mind is capable of rational judgment.

The characterization of some moral judgments as being *heuristic* deserves similar scrutiny, which I will provide in chapter 9.4. For now, note that, in order to understand whether *heuristic* processes produce a given judgment, it matters how the factors to which moral judgment *actually responds* relate to whatever it is that moral judgment is *supposed to track* (which is, in turn, what judgments of moral relevance are about).

Apart from the issue of global scope, Kahane imputed a second problem to evolutionary debunking arguments, namely, that they presuppose the contentious view that mind-independent moral truth exists (objectivism). However, the discussion of Greene's position clarified that, contrary to Kahane's assumption, subjectivist debunking is possible. Moreover, if we acknowledge Street's Darwinian Dilemma for objectivists *and* find that evolutionary influences indeed pervade moral judgment, we should reject moral objectivism. Knowledge of the origins and workings of moral judgment might even provide reasons to believe that subjectivism is a more accurate account of morality that are *independent* of evolutionary considerations.

Given these various ways in which empirical science bears on the cogency of philosophical positions, chapter 8 summarizes central insights regarding the significance of emotion, intuition, reason, and evolution for moral judgment identified in part II of this thesis. Chapter 9 assesses the positions recapitulated here in the light of this summary, and offers further conjectures about future philosophical repercussions of moral psychology.

# 8 Summary of Moral-Psychological Theories

## 8.1 From Morality vs. Convention to Moral Foundations and Relational Models

The exploration of the psychology of first- and second-order moral judgments in part II of this dissertation began with a look at the moral/conventional distinction. According to this conceptual framework, moral and conventional rules differ primarily with respect to the combinations of *rule content* and *type of response to violations* typical of them: Moral rules prohibit harm, injustice, or the violation of rights; conventional rules concern other matters. Moral rules are universally valid and serious, while conventional rules have limited validity, and violations are less grave. This characterization served as a first pass at a definition of the morally relevant and the morally efficacious. However, it conflicts with empirical evidence: Sometimes, features of iudicanda that are unrelated to harm, justice, or rights affect moral judgment (i.e., judgments that fit the moral response pattern). Even actions that are not harmful, not unjust, and do not violate rights can trigger signature moral responses. On the other hand, harmful actions do not always elicit the full moral response pattern. Finally, some response patterns differ from the signature moral and conventional types. Thus, the moral/conventional framework proved inadequate as descriptive account of both the morally relevant and the morally efficacious.

Richard Shweder's 'big three' theory of the moral domain attempted to integrate the discovery that issues beyond harm, rights, and justice matter for moral judgment. He identified three 'discourses', the ethics of autonomy, of community, and of divinity. Compared to the moral/conventional framework, Shweder's theory offered a broader conception of what members of different cultures consider morally relevant, and of what affects their judgment. The ethics of autonomy aim at protecting individuals' freedom to satisfy their preferences through values like freedom of choice, freedom from harm, or equality. The ethics of community serve to protect the integrity of social entities like groups and hierarchies; norms pertain to the specific roles each individual has in upholding social order and involve values like duty, hierarchy, interdependency, loyalty, or sacrifice. The ethics of divinity presume the existence of souls, and protect a special relationship between these souls

and a divine or natural order; rules govern action according to their effects on this relationship and express values like purity, sanctity, cleanliness, or a sacred order.[855] Shweder's theory was a significant advance in capturing the moral domain as understood also beyond Western societies, but it did not say much about the *origins* of the three 'ethics'.

Haidt and Joseph's moral foundations theory refined Shweder's model by further dividing the moral domain into six matters of fundamental concern: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, sanctity/degradation, and liberty/oppression. MFT holds that if we consider something morally relevant, it touches on at least one of these concerns. Apart from delineating the moral domain, Haidt and Joseph aim to explain *why* humans display these specific concerns, rather than others. Their approach is evolutionary: Each fundamental moral concern developed because, by motivating behavior and the establishment of corresponding norms, it helped abate or solve problems that individuals living in groups regularly encounter. Emotions play a pivotal role: They reliably connect certain events or circumstances in the environment to specific behaviors that tended to increase fitness. For instance, the anger felt by a victim of betrayal, or more precisely the actions caused by this emotion, on average made breaches of trust costlier for the perpetrator, thereby reducing their occurrence and increasing the fitness of both (reasonably) anger-prone and trustworthy agents. While the concrete norms developing from, as well as the relative weight of, the foundational concerns differ across cultures, individuals, and situations, the origin of moral judgment lies in the evolutionary processes that shaped humans and their ancestors into social animals.

Relational models theory provides a complementary perspective on morality: In RMT, how individuals judge depends on the relational model(s) they apply to a given social relationship. A specific *motive* governs each relational model, and the application of this motive determines which aspects of a iudicandum appear morally relevant. For instance, in communal-sharing relations, what matters is extent to which an action contributes to or detracts from the motive of unity. Whence do these models originate? For communal sharing and authority ranking relations, but not market pricing, Alan Fiske presumes predecessors in other species, the status of equality matching unclear.[856] However, even if market pricing and equality matching relations do not occur in nonhuman species, it is still possible that they evolved during the 5 to 7 million years since the developmental trajectories of Homo

[855] Values extracted from Shweder & Menon (2014), p. 358.
[856] See Fiske (n.d.).

sapiens and our closest relatives separated. Again, emotions are crucial: They assess relational potential, and motivate to establish and regulate relationships. Insofar as emotions link specific elicitors to specific action tendencies because these connections proved evolutionarily advantageous, the reliance of relational models on emotions points to an evolutionary origin. This does not imply that cultural or individual learning do not also shape emotion triggers and expression. However, similar phenomena in young children and related species suggest a significant influence of evolution.

In relational models theory, but particularly in moral foundations theory, emotions are central to the formation of the moral domain and the determination of moral judgment. Therefore, every claim about the evolution of the categories of moral relevance contained in these theories is also a claim about the evolution of these emotions. Both the extent to which emotions determine moral judgment, and the extent to which evolutionary processes shaped moral judgment, are focal points of the philosophical engagement with moral psychology. A thorough investigation of the relations between emotions and morality, and their evolutionary roots, is thus a logical next step.

## 8.2 Morality and Emotions

Emotions play a pivotal role in the understanding of morality that emerges from experiments such as those conducted by Greene and Haidt. According to moral foundations theory, emotional responses determine what we consider morally relevant. Thus, chapter 4 provided a detailed analysis of various emotions, their relation to morality, and their evolutionary history. Emotions are combinations of specific physiological changes, a certain 'feel', facial or motor expressions, action tendencies, and cognitive processes. I assume that emotions can contain so-called 'appraisals', i.e., microjudgments. This claim, in combination with the position that moral judgments contain emotions, generates a question that has to be addressed prior to further discussion of the function of emotions: If moral properties and concepts involve emotions, can these emotions in turn contain (moral) judgments? Even if emotions sometimes occur without appraisals, an evolutionary perspective suggests that emotions and appraisals function as joint entities in the context of morality: The ability to form appraisals can evolve only if it generates systematic fitness benefits, and it has no such benefits unless it affects survival or reproduction via, for instance, action tendencies contained in emotions.

Like other attempts to explain moral properties in terms of nonmoral properties, the claim that moral judgments are somehow equivalent to psychological phenomena is bound

to encounter open-question-type challenges. If questions like "Is moral property Y *really* just the presence of nonmoral fact X?" appear *open*, does this prove that moral property Y cannot consist in the 'presence of X'? I do not think so. Firstly, the question might be less likely to appear open with respect to comprehensive descriptions of emotions that include their experiential components. More importantly, the aptness of a proposed definition of a moral property to encounter doubt is *not* a good indicator for the validity of the definition. Psychological or other concepts that may take the place of X are typically *not the kind of input* processed by those mental mechanisms that can make something appear morally relevant. If the X proposed does not appear morally relevant, the definition will not be completely convincing *prima facie*, which gives rise to the impression that 'the question is open'. I suspect that failure of definitions or explanations to elicit impressions of moral relevance is an important reason why science and morality are often seen as separate realms.

Having argued that emotions *can* contain appraisals, and that open-question arguments do not disqualify sentimentalist accounts of morality, I discussed how emotions relate to morality. I proposed that emotions could have *foundational* and *instrumental* relations to morality: *Foundationally moral* emotions mark factors as morally relevant by linking them with phenomenologically specific experiences and particular action tendencies. To some extent, these responses to relatively salient features of iudicanda also determine moral judgment. In such cases, the morally relevant and the morally efficacious coincide. The foundational function of emotions determines the scope of the moral domain and thereby shapes moral norms.

On the other hand, emotions can be *instrumentally moral*. In this capacity, emotions promote morally commendable outcomes. Identifying instrumentally moral effects of emotions requires a moral code as a criterion (it does *not* require that the agent whose emotion we categorize be aware of this particular or any other moral code). Among instrumentally moral effects of emotions, further distinctions suggest themselves: Morally commendable outcomes can be intended or unintended. An intention to bring about morally commendable outcome X can be based on 'moral' motivation, i.e., the agent believes that morality demands X, or it could spring from nonmoral motivation, if the agent is unaware that the outcome he intends to bring about is what morality requires. Instrumentally moral effects of both moral and nonmoral motivation can be motivated intrinsically or extrinsically. Intrinsically morally motivated agents bring about morally commendable outcomes because knowingly conforming to the moral code is in itself rewarding. Extrinsically morally motivated agents bring about morally commendable outcomes because they expect some *other*

reward from conforming to the moral code, such as good fortune, social prestige, or a pleasant afterlife. Intrinsically, nonmorally motivated agents bring about morally commendable outcomes because they enjoy bringing about these outcomes regardless of the outcome's moral status. Extrinsically, nonmorally motivated agents bring about morally commendable outcomes because they expect some external reward for bringing about these outcomes, such as a finder's fee, where that reward is, at least in the agent's eyes, independent of the moral status of the outcome.

Following this classificatory effort, I discussed other- and self-conscious emotions, as well as empathy and sympathy, in terms of the proposed criteria and with respect to their phylogenesis. These emotions greatly affect interaction with others based on evaluations of our own actions, their actions, or predictions regarding their reactions. Anger, contempt, disgust, and jealousy are other-critical, while gratitude and elevation/awe are other-praising. Empathy and sympathy comprise several sensitivities for the predicament of others. Guilt, shame and embarrassment are important self-critical emotions, while pride is self-praising.

## 8.3 Models of Moral Judgment

Chapter 5 investigated the complex interaction of emotional processes, intuition, and reasoning in moral judgment. Apart from being of descriptive interest, these relations matter because not only evolutionary explicability and emotional influences, but also the allegedly *intuitive* character of moral judgments has raised doubts regarding their reliability. Singer and Greene argue that reason, rather than emotional intuition, is often necessary for adequate judgment. In order to evaluate such claims, an understanding of these psychological concepts and their interaction is required. It is widely believed that humans dispose of two distinct information processing systems, one quick, automatic, and largely unconscious (system 1), the other relatively slow, effortful, and conscious (system 2). Models of moral cognition differ with regard to the roles they attribute to these systems. In this thesis, moral judgment is understood along the lines of Haidt's social-intuitionist model (SIM), amended by ideas of other authors. According to the SIM, affective, intuitive processes (system 1) and social influences determine moral judgments; individual reasoning alters them only rarely. Greene remarks that characteristics of the iudicandum influence which system dominates, and that reason might be more important than Haidt suspects at least in processes of moral change, i.e., when substantial transformations occur in the moral outlook of individuals. He also rightly points out that social influences can generate *counterintuitive* judg-

ment, whereas in Haidt's SIM, socially triggered changes in judgment proceed via the elicitation of intuitions. Emotions are considered important for the onto- and phylogenetic development of moral judgment, while individual judgments differ in the degree of emotional activation involved. Another position, known under the headings of 'linguistic analogy' or 'moral grammar', claims that emotional evaluative responses to iudicanda are usually *preceded* by system-1-dominated, unconscious appraisals of the situation. However, the strict separation of appraisals and emotions advocated by this position seems implausible from a developmental perspective: The evolution of appraisal processes can be explained only by reference to the fitness benefits of behavioral tendencies contained in the emotional responses to which these appraisals are coupled. A close look at the connection between emotions and moral judgments reveals that changes in emotional activation often engender differences in moral judgments even if appraisals remain constant, indicating that emotions are sufficient at least for some moral judgments. Moreover, individual emotions have various triggers, which, at least to some extent, explains why a wide range of factors can affect moral judgment. Research on psychopaths suggests that emotions are necessary for normal moral judgment, and that moral judgment usually motivates corresponding behavior. Again, without behavioral consequences that require such motivation, it is hard to explain the evolutionary development of moral judgment to which the regular involvement of emotions points. While intuitions are important in moral judgment, evolutionary processes did not shape all intuitions to the same extent. Reasoning can engender new intuitions, for instance by way of deliberate appraisal shifts or choice of environment. New intuitions can also develop via implicit (unconscious) learning during ontogenesis; such intuitions may even generate better decisions than system 2 processing in complex circumstances. Conscious comparison of judgments regarding similar cases can lead to long-term change in intuitions regarding these cases. In these comparisons, the moral relevance of psychologically efficacious differences between the iudicanda under consideration is evaluated intuitively. I argue that, if these differences appear morally irrelevant, it is due to the *inactivity* of emotional processes whose activation is required for impressions of moral relevance.

## 8.4 Modularity, Innateness, and Disagreement

Chapter 6.1 discussed research that suggests highly specialized psychological mechanisms for social exchange in humans. Individuals reason much better about if-P-then-Q rules if these rules express a social contract involving a benefit and a requirement. We seem to dispose of a dedicated capacity for cheater detection, which presumably evolved to avoid

being exploited. If a similar degree of modularization characterizes moral cognition more generally, then the issues we consider moral are probably processed by a multitude of mechanisms evolved to solve various adaptive problems. Accordingly, the mechanisms incorporate rather diverse standards of desirability, manifesting in emotional responses of positive or negative valence. Such modularity of human cognition might conflict with philosophical endeavors to formulate a minimal number of principles that gauge or establish the moral status of *every* iudicandum. At the very least, it explains why such principles regularly generate judgments that clash with *some* intuitions. The idea of a modular mind is often accompanied by the notion that these modules evolved, and that the corresponding faculties are to some significant extent innate. Chapter 6.2 took a step back and reviewed doubts regarding the innateness of morality. Firstly, on some interpretations of the innateness claim, we should expect to find moral rules common to all cultures. Prinz argues that salient candidate rules are not universal. For instance, harming is not condemned universally. However, this observation counts as evidence only against the innateness of moral views that contain just such a universal disapprobation. Other regularities regarding the evaluation of harming could nevertheless be innate, for instance depending on social context or relational models. The fact that very young children already show a dislike for individuals who harm innocent agents, but a preference for those who harm wrongdoers, is a case in point. Nor does extensive individual learning about moral norms show that no significant aspect of morality is innate. The kind of module proposed by Haidt and colleagues, for instance, manifests in an ability to *learn* certain moral rules more easily than others. The evolutionary account is compatible with large variation between individuals, for both genetic and cultural reasons as well as differences in individual experience. Is a moral preference for equal division of resources innate? Prinz denies this: (Equal) sharing with kin might have evolved, but is, in his view, not moral. Sharing with nonkin, on the other hand, might be moral, but varies widely across cultures and thus does not seem to be an innate norm. For reasons given below, I disagree with Prinz's categorization of sharing with kin as nonmoral. Regarding sharing with nonkin, the mere presence of cultural variation does not show that there is no innate concern for resource division according to *some* rule (depending, for instance, on position in a hierarchy). Is a preference for reciprocity innate? Prinz claims that the experience of guilt when we do not reciprocate cannot have evolved, because competitors unburdened by such qualms would have crowded out guilty-minded individuals. I do not believe this is necessarily the case: Guilt and its anticipation protect agents from being punished as cheaters. Generally, innateness of significant parts of morality does not preclude substantial

cultural influence. Prinz also questions the evidence for a specialized cheater-detection module. He suggests an alternative, learning-based explanation for the much higher success rate subjects achieve in Wason selection tasks that involve a social contract. However, his explanation fails to account for the difference in performance on social-contract- and precautionary-rule reasoning observed in some individuals. Prinz furthermore argues that people with deficient moral judgment show deficits also in the execution of other mental functions, and that therefore the impaired mechanisms are not exclusive to *moral* functioning. I do not believe that nativism implies that they have to be. Individual cognitive mechanisms can be involved in several specialized modules. On the other hand, the fact that specific functions are localized in specific brain regions is also compatible with the acquisition of moral cognition during ontogenesis. Yet this only reminds us that nativism does not necessarily posit innate moral *principles*. Inherited propensities to adopt certain principles more readily than others are also conceivable.

Compared to the universal presence of moral rules in human societies, which is compatible with both nativist and nurturist accounts of morality, similarities between morality-related behaviors of humans and the behavior of nonhuman primates or other social species, whose capacity to learn is presumably much more limited, provides stronger support for innate aspects of morality. Prinz claims that the dissimilarities between the behavior of great apes and humans clearly outweigh the similarities. More importantly, even behavior that seems similar to human morality is not *moral* in other animals, because it is not done for the right reasons. Non-human animals are incapable of entertaining a notion of morality, and consequently unable to act from specifically *moral* reasons. In order to assess these claims, I listed morality-related behaviors and characteristics that occur also in nonhuman animals, and those that seem to be unique to humans. In nonhuman animals, researchers have found sharing of resources, helping, reciprocal exchange, reconciliation, as well as guilt-like expressions and behavior. There is evidence of problem solving, retributive behavior, consolation, conflict mediation, mother-child bonding, special treatment of the injured or disabled, awareness of group membership, attempts at deception and their discovery, emotional contagion, violence inhibition, incest avoidance, awareness of one's reputation, aversion to negative inequity, and some altruistic punishment. Mechanisms that are likely predecessors of emotions such as fear, anger, contempt, disgust, shame and embarrassment, guilt, pride, gratitude, and empathy exist in nonhuman primates. Certain relations in animal societies appear to be structured according to the communal-sharing and author-

ity-ranking models suggested by RMT, possibly also according to equality matching. Apparently, uniquely human morality-related behaviors include widespread third-party punishment, frequent reciprocal interaction with nonkin, helping nonkin without reciprocation, and socio-moral disgust. Chimps apparently care mainly about those they interact with, while humans are also concerned with third parties. Humans can imagine a future that differs substantially from the present, think about the mental states of others (theory of mind), and imitate interaction patterns more extensively. In contrast to other animals, which mostly act from a kind of intuition or impulse, humans can and often do ponder their actions. Prinz claims that only humans possess the concept of universal validity of rules and the emotion of guilt, whereas seemingly moral behavior in chimps and other nonhuman species is motivated mostly by fear or anger. I agree that if *moral* action requires an understanding of what you ought *morally* to do, then no nonhuman animal acts morally. However, I believe that in fact motivational mechanisms congenial to those of our primate relatives produce many human behaviors that we consider expressions of morality: Guilt, for instance, is unpleasant. Behavior in accordance with a moral code can result from a motivation to avoid this experience. According to Prinz's demanding criterion, such behavior is not moral, because it does not involve an intrinsic motivation to do what morality requires. It is also often unclear whether egoistic or altruistic desires ultimately motivate behavior that benefits others. Excluding behavior motivated by fear and awe from morality as Prinz does renders large proportions of behavior nonmoral that can plausibly count as part of morality. In my view, Prinz's notion is too narrow. Even in cases where such narrowly moral motivation is conceivable, it is often not the *cause* of action. In addition, even if we accepted Prinz's notion, the relevant universal moral rules are nevertheless shaped by evolved, emotional sensibilities for fundamental concerns. I employ a broader notion of morality. My subject matter is the ubiquitous human habit of evaluating how iudicanda relate to standards that regulate social existence. Therefore, more psychological mechanisms are relevant for the explanation of morality-related behavior, and more continuity exists between nonhuman animals and humans than from Prinz's point of view. Prinz argues that moral rules are universally valid, and that nonhuman animals have no such concept. Moral-psychological research has shown, however, that not all moral rules are considered universally valid. Possibly, many rules appeared to be 'universally valid' as long as humans lived in small bands of hunter-gatherers or small agricultural communities, and the contemporary notion of universality could be owed to awareness of and interaction with a much larger number and variety of

conspecifics. A chimpanzee alpha male might well expect that *all* conspecifics known to him respect his rank: Even the impression of universal validity could have evolved.

The remaining important characteristic of human morality is its reliance on the notion of a moral code. In my view, this feature likewise does not justify the claim that the dissimilarities between nonhuman animal behavior and human moral behavior outweigh the similarities. In particular, evolved emotional capacities confine the *content* of moral norms by establishing what can appear to be morally relevant. Were there a being equipped with the (proto)emotional life of a chimp and human reasoning capacity, it might well develop the concept of a moral code. In contrast, it is hardly possible to imagine the morality of a rational being without emotions, because it is unclear where its fundamental categories of value could originate. A last argument Prinz advances against evolutionary-psychological accounts of morality stems from philosophy of science: Ceteris paribus, more parsimonious theories, i.e., theories that posit fewer entities, are superior. According to Prinz, conceiving of morality as culturally transmitted is more parsimonious than nativist accounts since it assumes only a general learning capacity and cultural evolution, whereas nativism posits numerous evolved mental modules. My rejoinder is that both accounts are *not* equal in other relevant aspects: The evolutionary perspective appears to be more accurate and more fruitful as a description of reality; these advantages justify a more populous psychological ontology. Moral foundations theory, for instance, is a more comprehensive account of the domain of morality than the moral/conventional framework.

Based on the content of the previous chapters, 6.3 explored possible causes of moral disagreement. Individual judgments are the product of complex interactions between inherited traits and environmental influences. For instance, gene expression depends on environment, i.e., a specific genotype can generate distinct phenotypes in different environments. Genes can code for psychological mechanisms that are context sensitive (e.g., the ability to acquire the language spoken in one's childhood environment) or which make us seek out experiential input that calibrates other psychological mechanisms. Responsiveness to each of the six moral foundations is at least partly inherited, and presumably distributed according to a Bell curve. Nevertheless, information processing and interpretation also depends on individual experience. Since each individual's set of experiences is unique, even genetically identical monozygotic multiples process input in unique ways. Usually, individuals differ both in their genetic makeup and in their set of experiences. A focus on diversity helps to explain moral disagreement, and to see why typical normative accounts of morality fail to convince all individuals equally. Both genes and experience can cause differences in

the concretion of fundamental moral concerns or in the application of relational models, and these differences in turn can explain variations both in first-order- and relevance judgments. Since fundamental concerns are engraved in brain tissue and its electrochemical properties, whatever affects the structures, states, or processes involved, both experiential and genetic, can affect these concerns. Systematic differences between the brains of conservatives and liberals support this assumption. Moral disagreement can result both from differences in the weighting or the concretion of fundamental moral concerns. The more difficult it is to relate a given factor to these fundamental concerns, the more likely it is that it will be uncontroversially irrelevant.

# 9 Assessing Normative and Metaethical Significance

My aim in this chapter is twofold: I want to evaluate the philosophical responses to moral-psychological research presented in chapter 2 in light of the understanding of morality and moral judgment developed in part II, and arrive at some further conjectures regarding the moral-philosophical significance of this understanding.

Most of the philosophical positions I have presented draw conclusions from the discovery of specific features of at least some moral judgments. In particular, they address the evolutionary origin as well as the emotional and intuitive character of moral judgment, the potentially heuristic modus operandi of moral intuitions, and the susceptibility of moral judgment to the influence of morally irrelevant factors or differences. These motifs are sometimes conflated. I proceed by considering how the extent to which evolution, emotion, and intuition shape moral judgments affects arguments that doubt their adequacy. I argue that these three characteristics (shaped by evolutionary processes, emotional, intuitive) are virtually ubiquitous in moral judgment. Corresponding debunking arguments therefore have global scope, and the absurd consequences of their application strongly suggest that they are misguided. Moreover, inconsistency looms if the alleged debunking of evolved, emotional, and intuitive judgments rests on impressions of moral irrelevance that *share* the very features the debunking argument incriminates. In spelling out this line of thought below, I emphasize the nature and significance of assessments of moral relevance, as well as their dependence on the terminology typical of different levels of explanation.

I furthermore discuss whether moral intuitions are heuristics. To that end, I draw on the typology of relations between emotions and morality developed in chapter 4. If moral intuitions are heuristics, doubts regarding their reliability are in order, since heuristics work satisfactorily only under specific conditions. I have not elaborated on this issue in chapter 2 because it is more fruitfully explained and discussed using ideas introduced only in part II. Moreover, criticism of moral judgments that invokes their alleged heuristic character differs from references to emotion, intuition, or evolution in a way that warrants special treatment: *Categorizing* intuitions as moral heuristics depends on assumptions regarding the epistemic goal of moral judgment, as well as assumptions about how specific psychological processes relate to that goal. Importantly, at least the former is itself a *moral-philosophical* assumption. In order to classify a judgment as intuitive or emotional, it suffices in contrast to consider the judgment process itself. Neither do claims about whether evolutionary processes affect a judgment presuppose a moral philosophical stance.

## 9.1 Notes on Normative Significance

We have come across several invocations of 'normative significance', most explicitly in the debate between Greene and Berker. According to Berker, moral psychology (neuroscience, to be precise) does not provide new *normative* claims. Rather, the normative premise in Greene's most important argument stems from an intuition, not from an experimental result. Greene, in his rejoinder, defends the view that moral psychology *is* normatively significant, since normative intuitions *in combination with* psychological findings generate new normative judgments, or undermine previously held normative convictions.

It is expedient to distinguish the different *loci* at which change in normative claims originates. Call the position that *some* normative statements depend on empirical facts 'modest ethical empiricism'. Corresponding 'modest normative significance' of empirical data seems uncontroversial. For consequentialists, the causal effects of alternative courses of action determine their moral status. Empirical facts are indispensable also for the application of deontological rules: We need to establish, for instance, whether some act is an act of lying (by comparing the content of a statement to the beliefs of the speaker), a killing, or a token of some other type of action regulated by the normative framework under consideration. Another way to describe modest normative significance is to say that the change in normative propositions affected by the data occurs on the level of the *descriptive premises* required for the derivation of normative conclusions (concrete, act- or situation-related judgments) from normative, i.e., value-defining premises. Given a stable notion of moral value, the moral status of a iudicandum depends on the extent to which it realizes the value(s) set by these normative premises, and the extent of realization can only be determined empirically.

What Greene proposes is more ambitious: He seeks to *eliminate* specific classes of *normative premises* from such moral syllogisms by providing debunking explanations of how we came to have the corresponding evaluative attitudes.[857] If the debunking were successful, we could say that the respective findings have *destructive* normative significance. I argue that moral psychology is unlikely to achieve destructive normative significance with regard to the fundamental moral concerns established by evolved emotional intuitions.

[857] He calls this "challenging somebody's values". Greene (2010), p. 9.

## 9.2 The Significance of Evolution, Emotion, and Intuition

Moral foundations theory, and the additional evidence for an evolutionary development of emotional responses discussed in part II, indicate that the influence of evolutionary processes extends further than Greene assumes not only in breadth, i.e., regarding the variety of first-order moral judgments concerned, but also, as it were, in depth: It affects assessments of moral relevance as well. Greene's argument is aimed at characteristically deontological judgments (e.g., that it is never permissible to kill one individual in order to save several others), which are supposedly generated by evolved intuitive-emotional processes. Characteristically consequentialist judgments (e.g., that one death is better than five deaths), in contrast, are supposedly produced by controlled, conscious reasoning. Is it plausible that controlled processing is free from the influences that allegedly incriminate characteristically deontological judgments?

In an article coauthored with Fiery Cushman and Liane Young, Greene suggests that the core of the *consequentialist* welfare principle, the notion that "harm is bad", itself has an affective basis.[858] Possibly, emotional responses provide the negative valence attached to harm on which controlled processing operates. Somewhat surprisingly, Cushman et al. do not explicitly discuss the extent to which this affective aversion to harm is owed to evolutionary processes. Instead, they explore two versions of this affective-origin hypothesis that reflect a distinction between two kinds of emotions introduced in chapter 2.2: *Alarm-like* emotions generate the impression of incommensurable value; they tend to circumvent reasoning and dominate rapid decisions, whereas *currency-like* emotions figure as offsettable weights in more deliberative processes. While Cushman et al. leave the question which of these provides the affective foundation of the welfare principle unanswered; Greene has argued elsewhere that consequentialist reasoning probably rests on currency-like emotions.[859] In light of his misgivings about evolutionary influence on moral judgment, this is understandable, since he also claims that evolutionary processes shaped alarm-like emotional responses.[860] If alarm-like emotions were at the base of consequentialist deliberations, Greene would have to explain why the 'nonmoral nature' of evolutionary processes does not render consequentialist reasoning just as unreliable as deontological intuitions. In fact, the same challenge can be raised if 'harm is bad' arises from currency-like emotions. Firstly, it is not clear whether these two kinds of emotions can really be distinguished, or whether, for instance,

858 See Cushman et al. (2010). Note that the discussion there is limited to judgments about physical harm.
859 See Greene (2008b), p. 41.
860 See Greene (2003), p. 489.

a model according to which emotions occur with varying intensity, up to the kind of dominant role in decisions Greene ascribes to alarm-like emotions, would be more adequate. Secondly, the burden of proof lies with those who claim that a currency-like emotional response to the effect that 'harm is bad' does *not* have evolutionary origins, since the advantageousness of such an evaluation at least with respect to self and kin is straightforward. Greene has to explain either why affective responses at the basis of consequentialist reasoning are not owed to 'nonmoral features of our evolutionary history', or else show why such a pedigree renders alarm-like emotions unreliable, but not currency-like emotions. It is worth noting that the evolutionary history of emotional intuitions featured prominently in *The Secret Joke of Kant's Soul*, but is demoted in more recent writings to one among several features (apart from being intuitive, emotional, and heuristic) that warrant suspicion in moral judgments.[861] Nevertheless, a response remains necessary. This does not mean that no response is conceivable. In his *Notes* on Berker's arguments, Greene argues that we should not rely on our 'automatic' moral reactions to problems which are 'fundamentally new', i.e., which are complex and to which these automatic responses have not been attuned.[862] In my view, it remains unclear how Greene can grant evolved emotional judgments authority with respect to moral problems that are *not* fundamentally new, since he claimed in *The Secret Joke of Kant's Soul* that the amoral nature of evolutionary development engenders, or at least makes likely, the influence of morally irrelevant factors on moral judgments, *regardless of problem type*.

In any case, I believe that granting special authority to consequentialist reasoning or the welfare principle based on the putative *absence* of evolutionary, emotional, or intuitive influence is a mistake.[863] Due to the prevalence of evolutionary influences, intuition, and emotion, debunking arguments regarding moral judgments about physical harm lead to the absurd conclusion that all psychologically normal judgments are unreliable. Moreover, the same holds for debunking of moral judgments beyond physical harm with reference to evolution, intuition, or emotion: If MFT, or a different evolutionary-psychological theory that attributes similar importance to emotion and intuition, adequately describes the origin of moral concerns also *beyond* physical harm, then this kind of debunking implies distrust in all judgments based on those concerns. There are several reasons to believe that not only judgments about physical harm have an evolved, emotional-intuitive basis: The intuitive judg-

861 See Greene (2010), p. 12.
862 See ibid., p. 23.
863 See Kauppinen (2014), p. 297 for a similar argument.

ments humans at least sometimes make require the ability to cognize morally relevant features of a iudicandum without system 2 processing. More effortful moral reasoning can be modeled as modifying the input to intuitive systems (e.g., iterated cycles through links 6, 1, and 2 in the social-intuitionist model), shifting the burden of proof to those who maintain that such deliberation does *not at all* fall back on evolved, emotional intuitions. Consider role-taking in moral judgment: Assessing the exact effects of an action on others might require conscious reasoning, but evolved emotional-intuitive processing is necessary to arrive at an evaluation. Moral reasoning requires impressions of moral relevance, which fundamental moral concerns established by emotional intuitions and the culture-specific concepts built upon them provide. Reasoned judgments based on criteria that *fail* to involve evolved emotional-intuitive processes are likely to appear detached from common-sense morality and remain behaviorally inert.

Apart from implying an absurdly wide scope for evolutionary debunking arguments, the pervasive influence of evolutionary processes, emotion, and intuition indicated by the research discussed in part II suggests that debunking arguments that depend on the *moral irrelevance of factors* whose effects on moral judgment are due only to emotion, intuition, or evolution are *inconsistent*. The degree to which a factor appears morally relevant depends, I propose, on its aptitude to excite those evolved emotional-intuitive mechanisms whose activity marks fundamental moral concerns. If so, the irrelevance judgment in Greene's argument is *itself* owed to evolved emotional-intuitive mechanisms, in the sense that impressions of irrelevance result from *failure* of the respective factor to activate them. For instance, judgments about the moral irrelevance of the kinds of things that govern evolutionary processes are probably *themselves* a result of how our evolved psychology works. In fact, typical impressions of irrelevance at *all* levels of explanation are presumably owed to evolution to some extent. Whether the impression in question concerns 'personal force' or inclusive fitness, the psychological mechanisms responsible for it are probably the same.

In his rejoinder to Berker, Greene argues that we should not trust intuitive responses in fundamentally new situations. Automatic responses, he argues, are the result of three types of trial-and-error learning processes: Individual experience, learning from others, or evolution.[864] Situations or problems count as fundamentally new when we have no trial-and-error experience with it in one of these senses. The automatic settings of moral cognition are the result of trial-and-error experience with specific types of problems. Greene emphasizes that some current problems are complex and new. Since these problems are very different from

[864] See Greene (2010), p. 23.

the problems that shaped our automatic responses, it is highly unlikely that automatic responses to them produce good solutions. This is the case, Greene argues, virtually *regardless* of our criterion for a good solution.

It can be reasonable to question intuitive evaluative responses, for instance if the *consequences* of specific behaviors *today* can differ substantially from what they were under the conditions that shaped them via evolutionary or cultural selection processes. In my view, Greene's concern has merit, but it is expressed in a misleading and ultimately mistaken way. Demanding that we should not rely on automatic responses *at all* in dealing with complex new problems goes too far. As I have tried to show, our fundamental categories of value result from automatic responses to a significant extent, and it is not at all clear where criteria for what it means to solve a problem well could originate if we ignored *all* input provided by these responses. It might be more appropriate to caution against intuitive assessments of the *methods* by which we try to bring about a state of affairs that increases the realization of values that spring from the fundamental moral concerns. Such assessments are, in part, intuitions about causal effects of certain kinds of actions, and Greene mentions plausible reasons why they could be unfit for circumstances that differ from the environments in which they developed.

## 9.3 Support for Mind-Dependence

What about the metaethical significance of moral-psychological research? Street's causal premise seems true: Our evaluative attitudes are saturated with evolutionary influences. I have claimed above that this fact renders evolutionary debunking implausible, since it would 'debunk' most or all of our moral judgments. However, the objectivist has to either debunk or explain how judgments shaped by evolution track moral truth. Such a tracking account is also implausible, but for different reasons: It is an inferior scientific theory.

Could objectivists challenge the first implausibility claim by questioning the rejection of universal moral misdirection? What kind of a judgment is it to find global moral skepticism implausible? It rests on the belief that it is not the case that most of our moral judgments are completely off-track. One worry about this argument *against* evolutionary debunking is that the rejection of global skepticism could *itself* rest on impressions of an evolutionary origin, say, because evolution has equipped us with such strong moral convictions that we simply cannot conceive that all of them should be mistaken.[865] If this were true, then the belief that it is not the case that most of our moral judgments are completely off-track might

[865] I thank Olivier Roy for pointing this out.

carry substantially less or even no authority. Supposing that a hypothetical evolutionary pedigree of a basic trust in moral judgments does *not* cast doubt on this conviction, as those who reject global skepticism would have to, seems to beg the question against the (hypothetical) global evolutionary debunker.[866] In a sense, it amounts to assuming what is to be shown, namely, that evolutionary influence does *not per se* undermine the trustworthiness of moral judgments. I concede that I have no decisive argument to convince those willing to accept that virtually every moral judgment is misguided. I point to the explanatory power and coherence with a scientific worldview that mark the nonobjectivist perspective suggested by evolutionary-psychological accounts of moral judgment, and bite this bullet.

To elaborate: Let us assume that, given the pervasive influence of evolution on our moral judgments, objectivism indeed has to take either of two implausible positions, and that it should therefore be rejected. What about other forms of evolutionary debunking? Greene's variant in particular rests *not* on a notion of mind-independent moral truth, but on *intuitions* about moral irrelevance. This indicates a nonobjectivist conception within which moral judgments can nevertheless be adequate or inadequate. Adequacy depends not on their correspondence with mind-independent moral facts, but on whether judgments correctly take into account what is morally relevant. What is morally relevant depends in turn on our intuitions. If, as I have argued, objectivist evolutionary debunking is implausible because it implies global skepticism, *non*objectivist evolutionary debunking is just as implausible. If the nonobjectivist holds that evolutionary processes are unrelated to mind-*dependent* moral truth, and accepts that evolutionary influence pervades our evaluative attitudes, then he has to hold that none of these evaluative attitudes is justified. Does the nonobjectivist face a problem analogous to the second horn of the objectivist's Darwinian Dilemma, namely that accounts of how evaluative attitudes track mind-*dependent* moral truth are inferior to alternative explanations? On the nonobjectivist view, evaluative attitudes and the evolutionary processes that shaped them *are* related. This relation, however, differs in a crucial way from the relations the objectivist can posit, to wit, in terms of the *direction of dependence* between evolutionary processes and evaluative attitudes:[867] On the objectivist view, moral truths are *prior to* evolutionary processes, which either track them or not. On the nonobjectivist account suggested here, in contrast, evolutionary processes come first. The adequacy of moral judgments *depends on* the evaluative attitudes we evolved to have. Moreover, while the objectivist tracking account competes with evolutionary explanations of evaluative attitudes

[866] Actual debunkers typically do not aim at global moral skepticism; Greene explicitly states that the debunking must stop somewhere. See Greene (2008b), p. 76.

[867] See Street (2006), p. 154.

that do *not* posit mind-independent moral truth, the nonobjectivist account of the relation between moral 'truth' and our evaluative attitudes is compatible with such explanations. Therefore, and because the account of morality developed in Part II has considerable explanatory power, endorsing moral-psychological research corroborates an understanding of morality as mind-dependent, and weakens objectivism.

## 9.4 Are Moral Intuitions Heuristics?

Apart from arguments referring to the evolutionary history or emotional-intuitive character of moral judgments, Greene and others have based doubts regarding the reliability of moral intuitions on the claim that they are *heuristics*.[868] In the context of decision making, heuristics are rules that ignore part of the relevant information in order to produce reasonably good judgments under specific constraints (temporal, cognitive, etc.).[869] Understanding moral intuitions as heuristics emphasizes the *dependence* of evolved abilities on the characteristics of the environment they are applied in.[870] If intuitions are heuristics, they will regularly misjudge in circumstances that differ significantly from those they are adjusted to with respect to the correlation between information inputs and the actual target of the judgment.

[868] See for instance Greene (2014b), p. 217.
[869] See Gigerenzer (2008), p. 7.
[870] See ibid., pp. 7–8.

### *9.4.1 Understanding Moral Intuitions as Heuristics*

In order to find out whether moral intuitions really are heuristics, I consider a pertinent account of moral intuitions published by Walter Sinnott-Armstrong, Fiery Cushman, and Liane Young. On their definition, a heuristic "generates a judgment about a relatively inaccessible attribute T of an object by assessing a relatively more accessible heuristic attribute H."[871] The substitution of H for T is unconscious. If heuristics are explained in evolutionary terms, the judgment about H should be a good enough approximation of T; otherwise, it is hard to explain how the mechanism developed.[872] Strong correlation between the presence of H and judgments about the presence of T, and the observation that occurrences of H silence counterevidence (evidence against T), speak in favor of the presence of heuristics. Do moral intuitions fit this description?

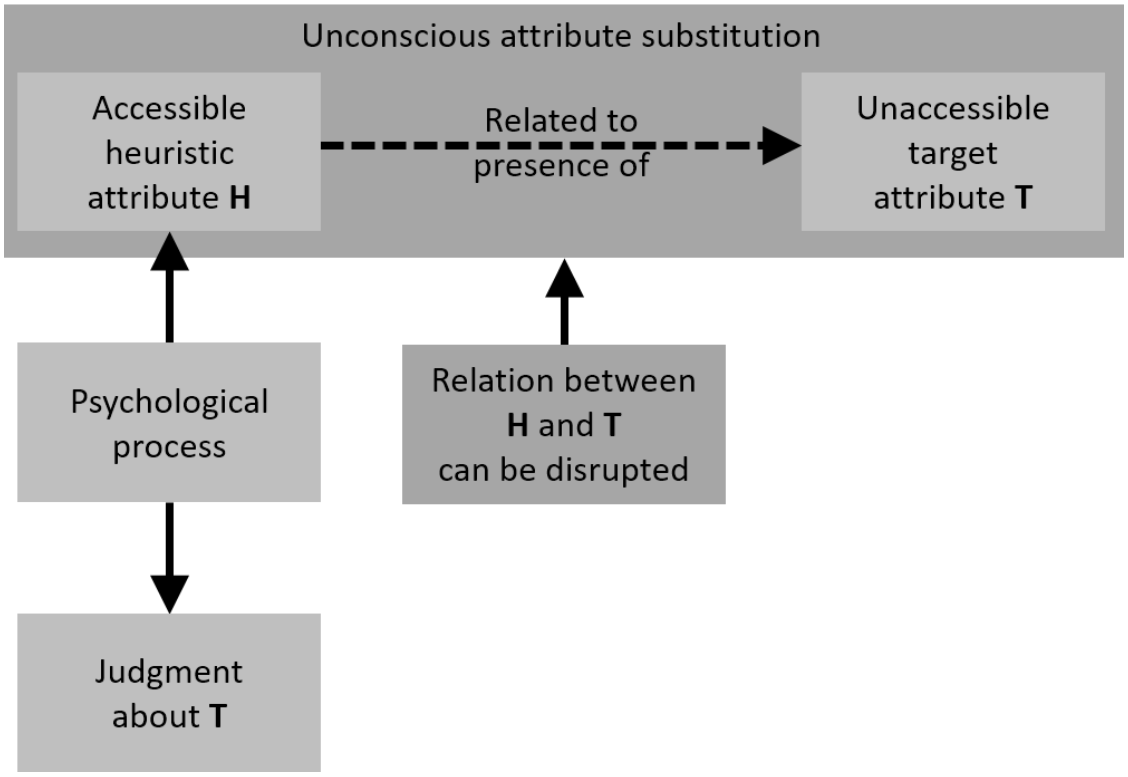


**Figure 7: The Heuristic Model of Moral Intuition**

Illustration by BH, based on Sinnott-Armstrong et al. (2010)

Moral intuitions concern the presence of a moral property T; Sinnott-Armstrong et al.'s example is *moral wrongness*. Is T relatively inaccessible? Upfront, Sinnott-Armstrong et al. claim that we cannot rely on *impressions* that T is accessible, since attribute substitution is

[871] See Sinnott-Armstrong et al. (2010), p. 251.

[872] I also assume that correctly assessing the presence of attribute T is evolutionarily advantageous.

supposedly unconscious. People might *believe* that they perceive T (and possibly quite easily), even though *in fact* they perceive H. How to proceed? The authors enumerate what they consider plausible notions of moral wrongness: consequentialist (does not produce the best consequences overall), Kantian (incompatible with the Categorical Imperative), contractarian (violates rules that all rational impartial people would accept [Rawls]; violates rules no reasonable person would reject [Scanlon]) and social moral relativist (violates conventions of a given society) accounts. None of these, they argue, is easily accessible. However, T would be easily accessible in case it consisted in the emotional state of the judge. Sinnott-Armstrong et al. dismiss this notion of moral wrongness as implausible, since it can ostensibly explain neither the occurrence of moral disagreement, nor "other common features of morality"[873]. Thus, since all plausible notions of T appear relatively inaccessible, they take intuitions about moral wrongness to satisfy the first part of the definition.

Next, Sinnott-Armstrong et al. attempt to determine the heuristic attribute H, and whether H is related to T, but more readily accessible. First, they discuss heuristics discovered in nonmoral contexts, such as "do what the majority does" or "I agree with people I like". Sinnott-Armstrong et al. concede that these may serve as H in some instances of moral judgment, but suspect that other heuristics specific to the domain of morality exist. Moreover, majorities and friends are sometimes criticized on moral grounds. The second class of candidate Hs are criteria mentioned in common-sense moral rules and principles, such as "do not kill". (In this case, the heuristic would be something like "If an action A is a killing/theft/lie/etc. (H), then A is morally wrong (T)"). This approach encounters several problems: Firstly, the criteria do not seem basic enough. For instance, not all killings/lies/etc. are considered wrong. Moreover, we process moral rules and principles consciously at least sometimes. There is also a methodological problem: If a property P is said to be a heuristic attribute H for a moral property T, the possibility that P *actually is* T should be eliminated. This, however, can only be achieved by making substantial assumptions about "what makes wrong acts wrong"[874]. For instance, some property of an act might appear as a heuristic attribute to a consequentialist, while to a deontologist, it is *itself* a wrong-making property. Such assumptions, however, are not part of a *scientific* evaluation of whether moral intuitions are heuristics or not, and would render resulting categorizations vulnerable to attacks from other moral-philosophical viewpoints. The same problem bedevils another class of possible heuristic attributes proposed by Cass R. Sunstein (the so-

873 Ibid., p. 257.
874 Ibid., p. 258.

called cold-heart heuristic, fee heuristic, betrayal heuristic, etc.)[875]: Again, it seems impossible to decide whether the properties suggested are Hs or Ts (in the sense of deontological rules) without committing to substantive *normative* assumptions. In addition, these 'heuristics', as well as moral rules and principles, are most often employed consciously. (We consciously think about whether some act amounts to a killing, etc.). Are there unconscious moral principles, i.e., patterns in moral judgment of which we are unaware? The aforementioned Doctrine of Double Effect might be an example. Research also points to further factors of whose influence we are unaware, such as whether someone is killed by the force of our muscles or not (personal force). Yet even if these factors do in fact unconsciously affect judgment, they are unlikely candidates for heuristic attributes because, according to Sinnott-Armstrong, they are not readily accessible (or not *more* readily accessible).

In the eyes of Sinnott-Armstrong et al., the most promising candidate is a general 'affect heuristic': A iudicandum X is condemned morally if thinking about X elicits (certain) unpleasant affective responses.[876] In contrast to other heuristics, the affect heuristic is not limited to specific types of actions. Sinnott-Armstrong et al. suggest that this mechanism could underlie many of the other candidate heuristics. In any case, negative affect is easily accessible, and presumably unconsciously substitutable. The question whether a given attribute is T or H does not arise, because Sinnott-Armstrong et al. have already dismissed the possibility that T consists in emotional states. The authors therefore continue exploring the affect heuristic. Are attributes substituted unconsciously, and is there a systematic relation between negative affect and moral wrongness? With respect to negative affect, statements of subjects about the presence of H strongly correlate with their statements about the presence of T, and manipulating affect changes moral judgment.[877] The substitution process seems to be unconscious: When people are confronted with the correlation between their moral judgments and their emotional states, they often deny any connection, which they probably would not do had they been aware of substituting negative affect for moral properties. Subjects also tend *not* to refer to their affective state when asked to explain their judgment.[878]

875 Cold-heart heuristic: "Those who know they will cause a death, and do so anyway, are regarded as cold-hearted monsters." Fee heuristic: "People should not be permitted to engage in moral wrongdoing for a fee." Betrayal heuristic: "Punish, and do not reward, betrayals of trust." See Sunstein (2005), pp. 536–537.

876 Sinnott-Armstrong et al. do not require that we experience negative affect *every* time an attribution of moral wrongness is made. What is required is that even generalized, nonaffective moral condemnation originates in former attributions of wrongness that (unconsciously) rest on negative affect.

877 See Sinnott-Armstrong et al. (2010), p. 263.

878 See ibid., pp. 266–267.

Finally, Sinnott-Armstrong and his coauthors consider philosophical consequences that might ensue if moral intuitions *are* heuristics. Firstly, moral intuitionists cannot claim *direct* insight into the presence of moral properties. Secondly, and most importantly for my purposes, if moral intuitions are indeed moral heuristics, they are probably unreliable in 'unusual' conditions. The authors admit, however, that it is currently quite unclear *which* moral intuitions are heuristics and *how* their reliability depends on the circumstances. Finally, they note that moral theories which posit relatively inaccessible notions of moral properties (e.g., "generating the best consequences overall") can, whenever their theory has counterintuitive implications, claim that the opposing *intuitions* are mistaken. However, as long as the moral theory in question counts correspondence with *other* intuitions among its virtues, it has to explain why *these* intuitions do a good job of approximating T.

### *9.4.2 Foundational Moral Intuitions are Not Heuristics*

I disagree with Sinnott-Armstrong et al. regarding the implications of the strong ties observed between emotional, intuitive mechanisms and moral judgment. In my view, moral intuitions that *establish* value are not heuristics. Let me explain. First, a remark on method. It is problematic to argue for the claim that moral intuitions are heuristics, but leave fundamental metaethical issues on which this categorization depends unaddressed. In particular, Sinnott-Armstrong and his coauthors raise no doubts regarding the evolutionary-psychological explicability of moral intuitions. In my view, supported by Street's argument, the evolutionary perspective makes a position of the kind Sinnott-Armstrong et al. refer to as 'skeptical' (i.e., that there is no moral reality beyond that established by certain intuitions) appear very plausible. If the skeptical view is correct, at least some moral intuitions are not heuristics.

In light of the relations between morality and emotions identified in chapter 4 and the models of moral cognition discussed in chapter 5, affect might very well be an important part of 'moral T' not just regarding moral wrongness, but most or even all moral properties. Sinnott-Armstrong et al.'s assertion that such an account cannot explain moral disagreement is not convincing. It presumably refers to familiar criticisms of early noncognitivist metaethical theories such as emotivism, according to which positions that base moral judgment in emotions cannot account for important features of moral debate, such as the fact that we *argue* about morality and do not treat it like a matter of taste.[879] However, as I have tried to show, an understanding of morality as based in emotional fundamental concerns is

[879] See Adler (2005), p. 543.

compatible with a lot of diversity in how these foundations are fleshed out by different individuals and cultures, and with perceiving one's own conviction as correct and binding also for others. The intuitive, emotional evaluations that identify fundamental concerns (subject to some degree of culture-dependent or idiosyncratic modification) are *not* heuristics. Without them, there would be no 'moral T' as we ordinarily envisage it. This is my primary disagreement with Sinnott-Armstrong et al. Nevertheless, my understanding of morality also has some use for the notion of heuristics. Recall, for instance, the mechanisms of intuition change described in chapter 5.5: Intuitive judgments can change due to deliberate appraisal shifts, willful exposition to other intuitions, transition of evaluative processing from system 2 to system 1, and result from implicit learning or consistency reasoning. On a continuum of psychological distance from fundamental moral concerns, the intuitions thus generated can be further removed than other intuitions. Consider an *intuitive*, positive affective response to a picture of Nelson Mandela as compared to an *intuitive*, positive affective response to a picture of a mother caressing her infant child. I reckon that the second response is not a heuristic, but a fundamental, evolved, emotional kernel of value, while the first response required system 2 activity in the judging individual's past to understand the Apartheid regime in South Africa, and can therefore count as heuristic. *Foundational moral intuitions are not heuristics.* Except in the context of discussions about moral enhancement, it makes little sense to question their adequacy, because they *establish* the fundamental categories of value. The central distinction of the different relations between emotions and morality made in chapter 4.3.5 comes in handy: When intuitions express the *foundationally moral* function of emotions, they are not heuristic. These psychological mechanisms constitute the moral properties about which we care.

In contrast, *heuristic* intuitions, i.e., those that manifest the *instrumentally moral* functions of emotions, can be assessed in terms of their aptitude to advance the realization of the values that spring from fundamental concerns. There is ample room for disagreement about this aptitude, not only due to diverging assumptions about the causal effects of specific evaluative habits, but also because the relative importance of the various moral foundations and their concretion can differ considerably across cultures and individuals. Heuristic intuitions are less trustworthy in unusual circumstances. Legitimate concerns about evolved, intuitive-emotional judgments appear to boil down to the kind of worry Greene expressed by means of his 'camera analogy'. In unusual conditions, we should not blindly rely on such judgments, but rather consciously take into account those environmental conditions that

affect the causal connections between iudicanda and the realization of the values that originate in the fundamental moral concerns. Those concerns, however, remain intact. In sum, the generalized distrust Sinnott-Armstrong et al. harbor towards intuitions is unwarranted. While some intuitions are heuristics, many are not, but are rather origins of moral values.

# 10 Concluding Thoughts

Many of the arguments analyzed in this thesis express a desire for cognitive convenience in the shape of clear distinctions, straightforward rules, and definite judgment in morality. Given the complexity of moral judgment, of which I am convinced we have only had a glimpse, this desire is inappropriate if one's aspiration is to develop a descriptively adequate understanding of reality. For the practical purpose of regulating social life on the other hand, simple rules and principles are without alternative, given the cognitive mechanics underlying human morality. Possibly, *normative ethics* can therefore never *both* provide guidelines for action that serve their purpose, *and* acknowledge the true complexity of moral experience. The *science of morality* in contrast will not, I predict, content itself with the relatively simple current models of moral judgment that still lend themselves to wholesale assessment of moral-philosophical schools of thought. Dual-process models, for instance, are a helpful simplification, but a simplification nevertheless. A descriptively adequate account of moral judgment will comprise many more distinctions of iudicanda, judgment contexts, cognitive processes, behavioral responses, etc. Mapping the trustworthiness of moral judgments on these distinctions is deeply problematic: Firstly, the sheer combinatorial complexity would make such an approach impractical. We cannot sift through hundreds of distinctions every time we want to assess or make a moral judgment. Secondly, the concepts that figure in scientific explanations are not the kind of input that can spark the motivation required for a moral code to fulfill its function. We are bound in this respect by the way our motivational capacities work.

The low aptitude for moral relevance of the explanatory concepts that the various scientific disciplines concerned with morality use in accordance with their respective levels of explanation guarantees that the scientific investigation of morality and moral judgment will continue to generate impressions of bias and error in moral judgment. Short of substantially altering the psychological mechanisms whose workings establish fundamental moral concerns, not much can be done to prevent these impressions. We should, however, be clear in our minds about the fact that *every* imaginable moral judgment can be explained in ways that generate the impression that it is affected by morally irrelevant factors. The degree to which a given concept appears morally relevant depends on its aptitude to excite the emotional psychological mechanisms that mark fundamental moral concerns.

What does this observation imply with respect to relevance-based debunking arguments? Probably, evolutionary, and (some) psychological, explanations of moral evaluations are

prone to elicit impressions of irrelevance, because concepts like 'inclusive fitness' are not suitable input for the mental mechanisms whose activation conveys that something is morally relevant. However, not every discovery of influences that appear morally irrelevant is *psychologically* sufficient for the dismissal of the corresponding evaluative attitude. Consider the tendency to feel a stronger obligation to care about the well-being of one's children than about the well-being of unrelated children far away. Evolutionary-psychological explanations for that attitude are readily available. Yet even if we accept that the overwhelming emotional pull of our own offspring's well-being is the product of processes that respond to morally irrelevant factors, it seems unlikely that anyone would conclude that the fact that a child is their own is *not* morally relevant, or even adjust their behavior accordingly. If we find, however, that hypnotically induced disgust makes moral judgments more severe, and can even cause moral condemnation without apparent cause, evolutionary accounts of *why* disgust can have such an effect might support the suspicion that this variation in judgment is not tracking any morally relevant difference.[880] Perhaps, moral-psychological research can persuade us to discard specific judgments, if the respective factor does not elicit impressions of moral relevance on *any* level of explanation, or if such impressions are not vivid enough to counterbalance the impressions of moral irrelevance.

[880] See Wheatley & Haidt (2005).

# Bibliography


**Adler, M. D. (2005):** Cognitivism, Controversy, and Moral Heuristics. In: *Behavioral and Brain Sciences*, 28 (04), pp. 542–543.

**Allman, J. M. & Woodward, J. (2008):** What are Moral Intuitions and Why Should We Care About Them? A Neurobiological Perspective. In: *Philosophical Issues*, 18 (1), pp. 164–185.

**Appiah, K. A. (2008):** *Experiments in Ethics*. Cambridge, MA: Harvard University Press.

**Aquino, K., McFerran, B. & Laven, M. (2011):** Moral identity and the experience of moral elevation in response to acts of uncommon goodness. In: *Journal of Personality and Social Psychology*, 100 (4), pp. 703–718.

**Arnold, M. B. (1960):** *Emotion and Personality Vol. 1: Psychological Aspects*. New York, NY: Columbia University Press.

**Bandura, A. (1999):** Moral disengagement in the perpetration of inhumanities. In: *Personality and Social Psychology Review*, 3 (3), pp. 193–209.

**Batson, C. D. (2011):** What's Wrong with Morality? In: *Emotion Review*, 3 (3), pp. 230–236.

**Baumeister, R. F. (2005):** *The cultural animal: Human nature, meaning, and social life*. Oxford: Oxford University Press.

**Berker, S. (2009):** The Normative Insignificance of Neuroscience. In: *Philosophy & Public Affairs*, 37 (4), pp. 293–329.

**Bierhoff, H.-W. (2002):** *Prosocial Behaviour*. Hove: Psychology Press.

**Blair, R. J. (1995):** A cognitive developmental approach to morality: Investigating the psychopath. In: *Cognition*, 57 (1), pp. 1–29.

**Blair, R. J. (2007):** The amygdala and ventromedial prefrontal cortex in morality and psychopathy. In: *Trends in Cognitive Sciences*, 11 (9), pp. 387–392.

**Bloom, P. (2010):** How do morals change? In: *Nature*, 464 (7288), p. 490.

**Bloom, P. (2012):** Moral Nativism and Moral Psychology. In: Mikulincer, M. & Shaver, P. R. (eds.): *The social psychology of morality: Exploring the causes of good and evil*. Washington, DC: American Psychological Association, pp. 71–89.

**Boehm, C. (2012):** *Moral Origins: The Evolution of Virtue, Altruism, and Shame*. New York, NY: Basic Books.

**Brosnan, S. F. (2012):** Introduction to "Justice in Animals". In: *Social Justice Research*, 25 (2), pp. 109–121.

**Brosnan, S. F. (2014):** Precursors of Morality – Evidence for Moral Behaviors in Non-human Primates. In: Christen, M., Van Schaik, C. P., Fischer, J., Huppenbauer, M. & Tanner, C. (eds.): *Empirically Informed Ethics: Morality between Facts and Norms*. Cham: Springer, pp. 85–98.

**Buss, D. M. (2008):** *Evolutionary Psychology: The New Science of the Mind,* 3. ed. Boston, MA: Pearson Allyn and Bacon.

**Campbell, R. & Kumar, V. (2012):** Moral Reasoning on the Ground. In: *Ethics*, 122 (2), pp. 273–312.

**Cela-Conde, C. J. (2005):** Did Evolution Fix Human Values? In: Changeux, J.-P., Damasio, A. R., Singer, W. & Christen, Y. (eds.): *Neurobiology of Human Values*. Berlin: Springer, pp. 11–15.

**Chapman, H. A. & Anderson, A. K. (2011):** Varieties of Moral Emotional Experience. In: *Emotion Review*, 3 (3), pp. 255–257.

**Churchland, P. S. (2011):** *Braintrust: What neuroscience tells us about morality*. Princeton, NJ: Princeton University Press.

**Cosmides, L. & Tooby, J. (1992):** Cognitive Adaptations for Social Exchange. In: Barkow, J. H., Cosmides, L. & Tooby, J. (eds.): *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York, NY: Oxford University Press, pp. 163–228.

**Cosmides, L. & Tooby, J. (2002):** Knowing thyself: The evolutionary psychology of moral reasoning and moral sentiments. In: *Business, Science, and Ethics*, 4, pp. 91–127.

**Cosmides, L. & Tooby, J. (2008):** Can a general deontic logic capture the facts of human moral reasoning? How the mind interprets social exchange rules and detects cheaters. In: Sinnott-Armstrong, W. (ed.): *The Evolution of Morality: Adaptations and Innateness*. Cambridge, MA: MIT Press, pp. 53–119.

**Cushman, F. A. (2011):** Moral Emotions from the Frog's Eye View. In: *Emotion Review*, 3 (3), pp. 261–263.

**Cushman, F. A., Young, L. & Greene, J. D. (2010):** Multi-system Moral Psychology. In: Doris, J. M. (ed.): *The Moral Psychology Handbook*. Oxford: Oxford University Press, pp. 47–71.

**Cushman, F. A., Young, L. & Hauser, M. D. (2006):** The Role of Conscious Reasoning and Intuition in Moral Judgment. In: *Psychological Science*, 17 (12), pp. 1082–1089.

**Dalgleish, T. (2004):** The Emotional Brain. In: *Nature Reviews Neuroscience*, 5 (7), pp. 582–589.

**Damasio, A. R. (2005):** The Neurobiological Grounding of Human Values. In: Changeux, J.-P., Damasio, A. R., Singer, W. & Christen, Y. (eds.): *Neurobiology of Human Values*. Berlin: Springer, pp. 47–56.

**Darley, J. M. & Batson, C. D. (1973):** 'From Jerusalem to Jericho': A Study of Situational and Dispositional Variables in Helping Behavior. In: *Journal of Personality and Social Psychology*, 27 (1), pp. 100–108.

**Darwall, S. L. (1998):** *Philosophical Ethics*. Boulder, CO: Westview Press.

**Darwin, C. (1871):** *The Descent of Man and Selection in Relation to Sex,* 2. ed. London: Murray, 1882.

**De Hooge, I. E., Nelissen, R. M., Breugelmans, S. M. & Zeelenberg, M. (2011):** What is moral about guilt? Acting "prosocially" at the disadvantage of others. In: *Journal of Personality and Social Psychology*, 100 (3), pp. 462–473.

**De Waal, F. B. M. (2004):** Evolutionary Ethics, Aggression, and Violence: Lessons from Primate Research. In: *The Journal of Law, Medicine & Ethics*, 32 (1), pp. 18–23.

**De Waal, F. B. M. (2005):** Homo Homini Lupus? Morality, the Social Instincts, and our Fellow Primates. In: Changeux, J.-P., Damasio, A. R., Singer, W. & Christen, Y. (eds.): *Neurobiology of Human Values*. Berlin: Springer, pp. 17–35.

**De Waal, F. B. M. (2008):** Putting the Altruism Back into Altruism: The Evolution of Empathy. In: *Annual Review of Psychology*, 59 (1), pp. 279–300.

**De Waal, F. B. M. (2013):** *The Bonobo and the Atheist: In Search of Humanism Among the Primates*. New York, NY: W. W. Norton & Company.

**De Waal, F. B. M. (2014):** Natural normativity: The 'is' and 'ought' of animal behavior. In: *Behaviour*, 151, pp. 185–204.

**Decety, J. & Jackson, P. L. (2004):** The Functional Architecture of Human Empathy. In: *Behavioral and Cognitive Neuroscience Reviews*, 3 (2), pp. 71–100.

**DeSteno, D., Bartlett, M. Y., Baumann, J., Williams, L. A. & Dickens, L. (2010):** Gratitude as moral sentiment: Emotion-guided cooperation in economic exchange. In: *Emotion*, 10 (2), pp. 289–293.

**DeSteno, D., Valdesolo, P. & Bartlett, M. Y. (2006):** Jealousy and the Threatened Self: Getting to the Heart of the Green-Eyed Monster. In: *Journal of Personality and Social Psychology*, 91 (4), pp. 626–641.

**Doris, J. M. & Stich, S. (2005):** As a Matter of Fact: Empirical Perspectives on Ethics. In: Jackson, F. & Smith, M. (eds.): *The Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press, pp. 114–152.

**Downes, S. M. (2010):** Evolutionary Psychology. In: Zalta, E. N. (ed.): *The Stanford Encyclopedia of Philosophy*, Fall 2010 ed. Stanford, CA: Metaphysics Research Lab, Stanford University.

**Dunbar, R. I., Barrett, L., Lycett, J. E. & Dunbar, R. (2007):** *Evolutionary Psychology: A Beginner's Guide*. Oxford: Oneworld Publications.

**Dwyer, S., Huebner, B. & Hauser, M. D. (2010):** The Linguistic Analogy: Motivations, Results, and Speculations. In: *Topics in Cognitive Science*, 2 (3), pp. 486–510.

**Ekman, P. (2003):** *Emotions Revealed: Recognizing faces and feelings to improve communication and emotional life,* 2. ed. New York, NY: Owl Books, 2007.

**Ellsworth, P. C. & Scherer, K. R. (2003):** Appraisal Processes in Emotion. In: Davidson, R. J., Scherer, K. R. & Goldsmith, H. H. (eds.): *Handbook of Affective Sciences*. Oxford: Oxford University Press, pp. 572–595.

**Elqayam, S. & St. Evans, J. B. (2011):** Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking. In: *Behavioral and Brain Sciences*, 34 (5), pp. 233–248.

**Else-Quest, N. M., Higgins, A., Allison, C. & Morton, L. C. (2012):** Gender Differences in Self-Conscious Emotional Experience: A Meta-Analysis. In: *Psychological Bulletin*, 138 (5), pp. 947–981.

**Ermer, E. & Kiehl, K. A. (2010):** Psychopaths Are Impaired in Social Exchange and Precautionary Reasoning. In: *Psychological Science*, 21 (10), pp. 1399–1405.

**Evans, D. (2001):** *Emotion: The Science of Sentiment*. Oxford: Oxford University Press.

**Fehr, E. & Gächter, S. (2002):** Altruistic punishment in humans. In: *Nature*, 415 (6868), pp. 137–140.

**Fessler, D. M. (2010):** Cultural congruence between investigators and participants masks the unknown unknowns: Shame research as an example. In: *Behavioral and Brain Sciences*, 33 (2-3), p. 92.

**Fiske, A. P. (n.d.):** *Human Sociality*. http://www.sscnet.ucla.edu/anthro/faculty/fiske/relmodov.htm (accessed 12.06.2017).

**Foot, P. (1967):** The Problem of Abortion and the Doctrine of the Double Effect. In: *Oxford Review*, 5, pp. 5–15.

**Forgas, J. P. (2003):** Affective Influences on Attitudes and Judgments. In: Davidson, R. J., Scherer, K. R. & Goldsmith, H. H. (eds.): *Handbook of Affective Sciences.* Oxford: Oxford University Press, pp. 596–618.

**Futuyma, D. J. (2005):** *Evolution*. Sunderland, MA: Sinauer Associates.

**Gigerenzer, G. (2007):** *Bauchentscheidungen: Die Intelligenz des Unbewussten und die Macht der Intuition,* 8. ed. Munich: Bertelsmann.

**Gigerenzer, G. (2008):** Moral Intuition = Fast and Frugal Heuristics? In: Sinnott-Armstrong, W. (ed.): *The Cognitive Science of Morality: Intuition and Diversity.* Cambridge, MA: MIT Press, pp. 1–26.

**Giner-Sorolla, R., Kamau, C. W. & Castano, E. (2010):** Guilt and Shame Through Recipients' Eyes: The Moderating Effect of Blame. In: *Social Psychology*, 41 (2), pp. 88–92.

**Graham, J., Haidt, J. & Nosek, B. A. (2009):** Liberals and conservatives rely on different sets of moral foundations. In: *Journal of Personality and Social Psychology*, 96 (5), pp. 1029–1046.

**Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S. P. & Ditto, P. H. (2011):** Mapping the Moral Domain. In: *Journal of Personality and Social Psychology*, 101 (2), pp. 366–385.

**Greene, J. D. (2002):** *The Terrible, Horrible, No Good, Very Bad Truth About Morality and What To Do About It.* Doctoral Dissertation, Princeton University. Princeton, NJ.

**Greene, J. D. (2003):** From Neural 'is' to Moral 'ought': What are the Moral Implications of Neuroscientific Moral Psychology? In: *Nature Reviews Neuroscience*, 4 (10), pp. 847–850.

**Greene, J. D. (2005a):** Cognitive Neuroscience and the Structure of the Moral Mind. In: Carruthers, P., Laurence, S. & Stich, S. (eds.): *The Innate Mind: Structure and Contents.* Oxford: Oxford University Press, pp. 338–352.

**Greene, J. D. (2005b):** Emotion and Cognition in Moral Judgment: Evidence from Neuroimaging. In: Changeux, J.-P., Damasio, A. R., Singer, W. & Christen, Y. (eds.): *Neurobiology of Human Values.* Berlin: Springer, pp. 57–66.

**Greene, J. D. (2008a):** Reply to Mikhail and Timmons. In: Sinnott-Armstrong, W. (ed.): *The Neuroscience of Morality: Emotion, Brain Disorders, and Development.* Cambridge, MA: MIT Press, pp. 105–117.

**Greene, J. D. (2008b):** The Secret Joke of Kant's Soul. In: Sinnott-Armstrong, W. (ed.): *The Neuroscience of Morality: Emotion, Brain Disorders, and Development.* Cambridge, MA: MIT Press, pp. 35–79.

**Greene, J. D. (2009):** Dual-process morality and the personal/impersonal distinction: A reply to McGuire, Langdon, Coltheart, and Mackenzie. In: *Journal of Experimental Social Psychology*, 45 (3), pp. 581–584.

**Greene, J. D. (2010):** *Notes on 'The Normative Insignificance of Neuroscience' by Selim Berker.* https://joshgreene.squarespace.com/s/notes-on-berker.pdf (accessed 12.06.2017).

**Greene, J. D. (2014a):** Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics. In: *Ethics*, 124 (4), pp. 695–726.

**Greene, J. D. (2014b):** *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. London: Atlantic Books.

**Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E. & Cohen, J. D. (2009):** Pushing Moral Buttons: The Interaction Between Personal Force and Intention in Moral Judgment. In: *Cognition*, 111 (3), pp. 364–371.

**Greene, J. D. & Haidt, J. (2002):** How (and Where) Does Moral Judgment Work? In: *Trends in Cognitive Sciences*, 6 (12), pp. 517–523.

**Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E. & Cohen, J. D. (2008):** Cognitive load selectively interferes with utilitarian moral judgment. In: *Cognition*, 107 (3), pp. 1144–1154.

**Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M. & Cohen, J. D. (2001):** An fMRI Investigation of Emotional Engagement in Moral Judgment. In: *Science*, 293 (5537), pp. 2105–2108.

**Hagen, E. H. (2005):** Controversial Issues in Evolutionary Psychology. In: Buss, D. M. (ed.): *The handbook of evolutionary psychology*. Hoboken, NJ: Wiley, pp. 145–173.

**Haidt, J. (2001):** The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. In: *Psychological Review*, 108 (4), pp. 814–834.

**Haidt, J. (2003a):** Elevation and the Positive Psychology of Morality. In: Keyes, C. L. M. & Haidt, J. (eds.): *Flourishing: Positive psychology and the life well-lived*. Washington, DC: American Psychological Association, pp. 275–289.

**Haidt, J. (2003b):** The emotional dog does learn new tricks: A reply to Pizarro and Bloom. In: *Psychological Review*, 110 (1), pp. 197–198.

**Haidt, J. (2003c):** The Moral Emotions. In: Davidson, R. J., Scherer, K. R. & Goldsmith, H. H. (eds.): *Handbook of Affective Sciences*. Oxford: Oxford University Press, pp. 852–870.

**Haidt, J. (2008):** Morality. In: *Perspectives on Psychological Science*, 3 (1), pp. 65–72.

**Haidt, J. (2012):** *The Righteous Mind: Why Good People are Divided by Politics and Religion*. New York, NY: Pantheon Books.

**Haidt, J. & Bjorklund, F. (2008):** Social Intuitionists Answer Six Questions about Moral Psychology. In: Sinnott-Armstrong, W. (ed.): *The Cognitive Science of Morality: Intuition and Diversity*. Cambridge, MA: MIT Press, pp. 181–217.

**Haidt, J. & Graham, J. (2007):** When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. In: *Social Justice Research*, 20 (1), pp. 98–116.

**Haidt, J. & Hersh, M. A. (2001):** Sexual Morality: The Cultures and Emotions of Conservatives and Liberals. In: *Journal of Applied Social Psychology*, 31 (1), pp. 191–221.

**Haidt, J. & Joseph, C. (2004):** Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. In: *Daedalus*, 133 (4), pp. 55–66.

**Haidt, J. & Joseph, C. (2007):** The Moral Mind: How Five Sets of Innate Intuitions Guide the Development of Many Culture-Specific Virtues, and Perhaps Even Modules. In: Carruthers, P., Laurence, S. & Stich, S. (eds.): *The Innate Mind: Foundations and the Future*. New York, NY: Oxford University Press, pp. 367–391.

**Haidt, J. & Kesebir, S. (2010):** Morality. In: Fiske, S. T., Gilbert, D. T. & Lindzey, G. (eds.): *Handbook of Social Psychology*, 5. ed. Hoboken, NJ: Wiley, pp. 797–832.

**Haidt, J., Koller, S. H. & Dias, M. G. (1993):** Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog? In: *Journal of Personality and Social Psychology*, 65 (4), pp. 613–628.

**Haney, C., Banks, C. & Zimbardo, P. G. (1973):** Interpersonal Dynamics in a Simulated Prison. In: *International Journal of Criminology and Penology*, 1 (1), pp. 69–97.

**Harris, S. (2010):** *The moral landscape: How science can determine human values*. New York, NY: Free Press.

**Hauser, M. D. (2007):** *Moral Minds: The Nature of Right and Wrong*. New York, NY: HarperCollins e-books.

**Hauser, M. D., Young, L. & Cushman, F. A. (2008):** On Misreading the Linguistic Analogy: Response to Jesse Prinz and Ron Mallon. In: Sinnott-Armstrong, W. (ed.): *The Cognitive Science of Morality: Intuition and Diversity*. Cambridge, MA: MIT Press, pp. 171–179.

**Heidbrink, H. (2008):** *Einführung in die Moralpsychologie,* 3. ed. Weinheim: Beltz.

**Henrich, J., Heine, S. J. & Norenzayan, A. (2010):** The weirdest people in the world? In: *Behavioral and Brain Sciences*, 33 (2-3), pp. 61–83.

**Horberg, E. J., Oveis, C. & Keltner, D. (2011):** Emotions as Moral Amplifiers: An Appraisal Tendency Approach to the Influences of Distinct Emotions upon Moral Judgment. In: *Emotion Review*, 3 (3), pp. 237–244.

**Huebner, B., Dwyer, S. & Hauser, M. D. (2009):** The Role of Emotion in Moral Psychology. In: *Trends in Cognitive Sciences*, 13 (1), pp. 1–6.

**Huppert, B. (2010):** Sources of Moral Significance – How Evolved Emotions Inevitably Shape Moral Judgment. In: *Philosophy and Economics – Diskussionspapiere an der Universität Bayreuth*, 7 (16).

**Hutcherson, C. A. & Gross, J. J. (2011):** The moral emotions: A social–functionalist account of anger, disgust, and contempt. In: *Journal of Personality and Social Psychology*, 100 (4), pp. 719–737.

**Hüther, G. (2011):** *Bedienungsanleitung für ein menschliches Gehirn,* 9. ed. Göttingen: Vandenhoeck & Ruprecht.

**Hynes, C. A. (2008):** Morality, Inhibition, and Propositional Content. In: Sinnott-Armstrong, W. (ed.): *The Neuroscience of Morality: Emotion, Brain Disorders, and Development*. Cambridge, MA: MIT Press, pp. 25–30.

**Inbar, Y. & Pizarro, D. A. (2014):** Pollution and Purity in Moral and Political Judgment. In: Sarkissian, H. & Cole Wright, J. (eds.): *Advances in Experimental Moral Psychology*. London: Bloomsbury Publishing, pp. 111–129.

**Isen, A. M. & Levin, P. F. (1972):** Effect of Feeling Good on Helping: Cookies and Kindness. In: *Journal of Personality and Social Psychology*, 21 (3), pp. 384–388.

**James, S. M. (2011):** *An introduction to evolutionary ethics*. Malden, MA: Wiley-Blackwell.

**Jensen, N. H. & Petersen, M. B. (2011):** To Defer or To Stand Up? How Offender Formidability Affects Third Party Moral Outrage. In: *Evolutionary Psychology*, 9 (1), pp. 118–136.

**Jones, D. (2007):** Moral Psychology: The Depths of Disgust. In: *Nature*, 447 (7146), pp. 768–771.

**Kahane, G. (2011):** Evolutionary Debunking Arguments. In: *Noûs*, 45 (1), pp. 103–125.

**Kahane, G. & Shackel, N. (2010):** Methodological Issues in the Neuroscience of Moral Judgment. In: *Mind & Language*, 255 (5), pp. 561–582.

**Kahneman, D. & Sunstein, C. R. (2005):** Cognitive Psychology of Moral Intuitions. In: Changeux, J.-P., Damasio, A. R., Singer, W. & Christen, Y. (eds.): *Neurobiology of Human Values*. Berlin: Springer, pp. 91–105.

**Kanai, R., Feilden, T., Firth, C. & Rees, G. (2011):** Political orientations are correlated with brain structure in young adults. In: *Current Biology*, 21 (8), pp. 677–680.

**Kauppinen, A. (2014):** Ethics and Empirical Psychology – Critical Remarks to Empirically Informed Ethics. In: Christen, M., Van Schaik, C. P., Fischer, J., Huppenbauer, M. & Tanner, C. (eds.): *Empirically Informed Ethics: Morality between Facts and Norms*. Cham: Springer, pp. 279–305.

**Kelly, D. (2014):** Selective Debunking Arguments, Folk Psychology, and Empirical Moral Psychology. In: Sarkissian, H. & Cole Wright, J. (eds.): *Advances in Experimental Moral Psychology*. London: Bloomsbury Publishing, pp. 130–147.

**Kelly, D. R. & Stich, S. (2007):** Two Theories About the Cognitive Architecture Underlying Morality. In: Carruthers, P., Laurence, S. & Stich, S. (eds.): *The Innate Mind: Foundations and the Future*. New York, NY: Oxford University Press, pp. 348–366.

**Kennett, J. & Fine, C. (2009):** Will the Real Moral Judgment Please Stand Up? The Implications of Social Intuitionist Models of Cognition for Meta-ethics and Moral Psychology. In: *Ethical Theory and Moral Practice*, 12 (1), pp. 77–96.

**Kitcher, P. (1985):** *Vaulting Ambition: Sociobiology and the Quest for Human Nature*. Cambridge, MA: MIT Press.

**Kitcher, P. (2011):** *The Ethical Project*. Cambridge, MA: Harvard University Press.

**Knobe, J. (2003):** Intentional action and side effects in ordinary language. In: *Analysis*, 63 (279), pp. 190–194.

**Kohlberg, L. (1973):** The Claim to Moral Adequacy of a Highest Stage of Moral Judgment. In: *The Journal of Philosophy*, 70 (18), pp. 630–646.

**Kubacka, K. E., Finkenauer, C., Rusbult, C. E. & Keijsers, L. (2011):** Maintaining Close Relationships: Gratitude as a Motivator and a Detector of Maintenance Behavior. In: *Personality and Social Psychology Bulletin*, 37 (10), pp. 1362–1375.

**Kumar, V. & Campbell, R. (2012):** On the normative significance of experimental moral psychology. In: *Philosophical Psychology*, 25 (3), pp. 311–330.

**Lapsley, D. K. (1996):** *Moral psychology*. Boulder, CO: Westview Press.

**Lazarus, R. S. (1991):** *Emotion and adaptation*. New York, NY: Oxford University Press.

**Levy, N. (2007):** *Neuroethics: Challenges for the 21st Century*. Cambridge, MA: Cambridge University Press.

**Lewis, G. J. & Bates, T. C. (2011):** From left to right: How the personality system allows basic traits to influence politics via characteristic moral adaptations. In: *British Journal of Psychology*, 102 (3), pp. 546–558.

**Liu, C.-J. & Hao, F. (2011):** An Application of a Dual-Process Approach to Decision Making in Social Dilemmas. In: *American Journal of Psychology*, 124 (2), pp. 203–212.

**Loewenstein, G. & Lerner, J. S. (2003):** The Role of Affect in Decision Making. In: Davidson, R. J., Scherer, K. R. & Goldsmith, H. H. (eds.): *Handbook of Affective Sciences*. Oxford: Oxford University Press, pp. 619–642.

**Looren de Jong, H. (2011):** Evolutionary Psychology and Morality. Review Essay. In: *Ethical Theory and Moral Practice*, 14 (1), pp. 117–125.

**Mallon, R. (2008):** Ought We to Abandon a Domain-General Treatment of 'Ought'? In: Sinnott-Armstrong, W. (ed.): *The Evolution of Morality: Adaptations and Innateness*. Cambridge, MA: MIT Press, pp. 121–130.

**Mallon, R. & Nichols, S. (2011):** Dual Processes and Moral Rules. In: *Emotion Review*, 3 (3), pp. 284–285.

**Mascaro, S., Korb, K. B., Nicholson, A. E. & Woodberry, O. (2010):** *Evolving ethics*. Exeter: Imprint Academic.

**Mathews, K. E. & Canon, L. K. (1975):** Environmental Noise Level as a Determinant of Helping Behavior. In: *Journal of Personality and Social Psychology*, 32 (4), pp. 571–577.

**McGuire, J., Langdon, R., Coltheart, M. & Mackenzie, C. (2009):** A reanalysis of the personal/impersonal distinction in moral psychology research. In: *Journal of Experimental Social Psychology*, 45 (3), pp. 577–580.

**Milgram, S. (1963):** Behavioral Study of Obedience. In: *Journal of Abnormal and Social Psychology*, 67 (4), pp. 371–378.

**Moll, J., de Oliveira-Souza, R. & Zahn, R. (2008a):** The Neural Basis of Moral Cognition: Sentiments, Concepts, and Values. In: *Annals of the New York Academy of Sciences*, 1124 (1), pp. 161–180.

**Moll, J., de Oliveira-Souza, R., Zahn, R. & Grafman, J. (2008b):** The Cognitive Neuroscience of Moral Emotions. In: Sinnott-Armstrong, W. (ed.): *The Neuroscience of Morality: Emotion, Brain Disorders, and Development*. Cambridge, MA: MIT Press, pp. 1–17.

**Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F. & Grafman, J. (2005):** The Neural Basis of Human Moral Cognition. In: *Nature Reviews Neuroscience*, 6 (10), pp. 799–809.

**Nado, J., Kelly, D. R. & Stich, S. (2009):** Moral Judgment. In: Symons, J. & Calvo, P. (eds.): *The Routledge Companion to Philosophy of Psychology*. London: Routledge, pp. 621–633.

**Nichols, S. (2002):** Norms With Feeling: Towards a Psychological Account of Moral Judgment. In: *Cognition*, 84 (2), pp. 221–236.

**Nisan, M. (1987):** Moral Norms and Social Conventions: A Cross-Cultural Comparison. In: *Developmental Psychology*, 23 (5), pp. 719–725.

**Paxton, J. M. & Greene, J. D. (2010):** Moral Reasoning: Hints and Allegations. In: *Topics in Cognitive Science*, 2 (3), pp. 511–527.

**Pinker, S. (2001):** How the mind works. In: Appleman, P. (ed.): *Darwin: Texts commentary*, 3. ed. New York, NY: Norton, pp. 465–477.

**Pinker, S. (2006):** The Blank Slate. In: *The General Psychologist*, 41 (1), pp. 1–8.

**Pizarro, D. A. & Bloom, P. (2003):** The intelligence of the moral intuitions: A comment on Haidt (2001). In: *Psychological Review*, 110 (1), pp. 193–196.

**Pogge, T. (2005):** World poverty and human rights. In: *Ethics & International Affairs*, 19 (1), pp. 1–7.

**Preston, S. D. & De Waal, F. B. M. (2001):** Empathy: Its ultimate and proximate bases. In: *Behavioral and Brain Sciences*, 25 (1), pp. 1–20.

**Prinz, J. J. (2007a):** Can moral obligations be empirically discovered? In: *Midwest Studies in Philosophy*, 31 (1), pp. 271–291.

**Prinz, J. J. (2007b):** *The emotional construction of morals*. Oxford: Oxford University Press.

**Prinz, J. J. (2011):** Is Empathy Necessary for Morality? In: Coplan, A. & Goldie, P. (eds.): *Empathy: Philosophical and Psychological Perspectives*. Oxford: Oxford University Press, pp. 211–229.

**Prinz, J. J. (2012):** *Beyond Human Nature: How Culture and Experience Shape Our Lives*. London: Allen Lane.

**Prinz, J. J. (2014):** Where Do Morals Come From? – A Plea for a Cultural Approach. In: Christen, M., Van Schaik, C. P., Fischer, J., Huppenbauer, M. & Tanner, C. (eds.): *Empirically Informed Ethics: Morality between Facts and Norms*. Cham: Springer, pp. 99–116.

**Prinz, J. J. & Nichols, S. (2010):** Moral Emotions. In: Doris, J. M. (ed.): *The Moral Psychology Handbook*. Oxford: Oxford University Press, pp. 111–146.

**Rai, T. S. & Fiske, A. P. (2011):** Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. In: *Psychological Review*, 118 (1), pp. 57–75.

**Rochat, P. (2010):** What is really wrong with a priori claims of universality? Sampling, validity, process level, and the irresistible drive to reduce. In: *Behavioral and Brain Sciences*, 33 (2-3), pp. 107–108.

**Roeser, S. (2011):** *Moral emotions and intuitions*. Basingstoke: Palgrave Macmillan.

**Ross, L. & Nisbett, R. E. (1991):** The Person and the Situation. In: Nadelhoffer, T., Nahmias, E. A. & Nichols, S. (eds.): *The Person and the Situation*. Malden, MA: Wiley-Blackwell, 2010, pp. 187–196.

**Royzman, E. B., Leeman, R. F. & Baron, J. (2009):** Unsentimental ethics: Towards a content-specific account of the moral–conventional distinction. In: *Cognition*, 112 (1), pp. 159–174.

**Rozin, P., Haidt, J. & McCauley, C. R. (2008):** Disgust. In: Lewis, M., Haviland-Jones, J. M. & Barrett, L. F. (eds.): *Handbook of Emotions*, 3. ed. New York, NY: Guilford Press, pp. 757–776.

**Rozin, P., Lowery, L., Imada, S. & Haidt, J. (1999):** The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). In: *Journal of Personality and Social Psychology*, 76 (4), pp. 574–586.

**Sauer, H. (2012):** Morally irrelevant factors: What's left of the dual process-model of moral cognition? In: *Philosophical Psychology*, 25 (6), pp. 783–811.

**Schacter, D. L., Gilbert, D. T. & Wegner, D. M. (2009):** *Psychology*. New York, NY: Worth Publishers.

**Scherer, K. R. (2003):** Introduction: Cognitive Components of Emotion. In: Davidson, R. J., Scherer, K. R. & Goldsmith, H. H. (eds.): *Handbook of Affective Sciences*. Oxford: Oxford University Press, pp. 563–571.

**Schnall, S., Haidt, J., Clore, G. L. & Jordan, A. H. (2008):** Disgust as Embodied Moral Judgment. In: *Personality and Social Psychology Bulletin*, 34 (8), pp. 1096–1109.

**Shweder, R. A. (1990):** In Defense of Moral Realism: Reply to Gabennesch. In: *Child Development*, 61 (6), pp. 2060–2067.

**Shweder, R. A., Mahapatra, M. & Miller, J. G. (1990):** Culture and Moral Development. In: Kagan, J. & Lamb, S. (eds.): *The Emergence of Morality in Young Children*, 2. ed. Chicago: University of Chicago Press, pp. 1–82.

**Shweder, R. A. & Menon, U. (2014):** Old questions for the new anthropology of morality: A commentary. In: *Anthropological Theory*, 14 (3), pp. 356–370.

**Shweder, R. A., Much, N. C., Mahapatra, M. & Park, L. (2003):** The 'Big Three' of Morality (Autonomy, Community, Divinity) and the 'Big Three' Explanations of Suffering. In: Shweder, R. A. (ed.): *Why do Men Barbecue? Recipes for Cultural Psychology*. Cambridge, MA: Harvard University Press, pp. 119–166.

**Sidgwick, H. (1874):** *The Methods of Ethics,* 7. ed. London: Macmillan, 1907.

**Singer, P. (1972):** Famine, Affluence, and Morality. In: *Philosophy & Public Affairs*, 1 (3), pp. 229–243.

**Singer, P. (2005):** Ethics and Intuitions. In: *The Journal of Ethics*, 9 (3), pp. 331–352.

**Sinnott-Armstrong, W., Cushman, F. A. & Young, L. (2010):** Moral Intuitions. In: Doris, J. M. (ed.): *The Moral Psychology Handbook*. Oxford: Oxford University Press, pp. 246–272.

**Smith, E. E., Nolen-Hoeksema, S., Fredrickson, B. L. & Loftus, G. R. (2007):** *Atkinsons und Hilgards Einführung in die Psychologie,* 2. ed. Heidelberg: Spektrum.

**Sober, E. (1997):** Is the Mind an Adaptation for Coping with Environmental Complexity? In: *Biology & Philosophy*, 12 (4), pp. 539–550.

**Sober, E. (1998):** Prospects for an Evolutionary Ethics. In: Pojman, L. P. (ed.): *Ethical theory: Classical and contemporary readings*, 3. ed. Belmont, CA: Wadsworth, pp. 131–143.

**Sober, E. (2000):** *Philosophy of Biology,* 2. ed. Boulder, CO: Westview Press.

**Stich, S., Doris, J. M. & Roedder, E. (2010):** Altruism. In: Doris, J. M. (ed.): *The Moral Psychology Handbook*. Oxford: Oxford University Press, pp. 147–205.

**Street, S. (2006):** A Darwinian Dilemma for Realist Theories of Value. In: *Philosophical Studies*, 127 (1), pp. 109–166.

**Stueber, K. R. (2008):** Empathy. In: Zalta, E. N. (ed.): *The Stanford Encyclopedia of Philosophy*, Fall 2008 ed. Stanford, CA: Metaphysics Research Lab, Stanford University.

**Suhler, C. L. & Churchland, P. S. (2011):** Can Innate, Modular "Foundations" Explain Morality? Challenges for Haidt's Moral Foundations Theory. In: *Journal of Cognitive Neuroscience*, 23 (9), pp. 2103–2116.

**Sunar, D. (2009):** Suggestions for a New Integration in the Psychology of Morality. In: *Social and Personality Psychology Compass*, 3 (4), pp. 447–474.

**Sunstein, C. R. (2005):** Moral heuristics. In: *Behavioral and Brain Sciences*, 28 (04), pp. 531–542.

**Tangney, J. P., Stuewig, J. & Mashek, D. J. (2007):** Moral Emotions and Moral Behavior. In: *Annual Review of Psychology*, 58 (1), pp. 345–372.

**TenHouten, W. D. (2009):** *A general theory of emotions and social life*. London: Routledge.

**Thomson, J. J. (1985):** The Trolley Problem. In: *The Yale Law Journal*, 94 (6), pp. 1395–1415.

**Tooby, J. & Cosmides, L. (2001):** Does Beauty Build Adapted Minds? Toward an Evolutionary Theory of Aesthetics, Fiction, and the Arts. In: *SubStance*, 30 (1), pp. 6–27.

**Turiel, E. (1982):** Die Entwicklung sozial-konventionaler und moralischer Konzepte. In: Edelstein, W. & Keller, M. (eds.): *Perspektivität und Interpretation: Beiträge zur Entwicklung des sozialen Verstehens*. Frankfurt am Main: Suhrkamp Verlag, pp. 146–187.

**Turiel, E. (2006a):** The Development of Morality. In: Eisenberg, N., Damon, W. & Lerner, R. M. (eds.): *Social, emotional, and personality development*, 6. ed. Hoboken, NJ: Wiley, pp. 789–857.

**Turiel, E. (2006b):** Thought, Emotions, and Social Interactional Processes in Moral Development. In: Killen, M. & Smetana, J. G. (eds.): *Handbook of Moral Development*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 7–35.

**Valdesolo, P. & DeSteno, D. (2006):** Manipulations of Emotional Context Shape Moral Judgment. In: *Psychological Science*, 17 (6), pp. 476–477.

**Valdesolo, P. & DeSteno, D. (2011):** The Virtue in Vice: Short-Sightedness in the Study of Moral Emotions. In: *Emotion Review*, 3 (3), pp. 276–277.

**Van Schaik, C. P., Burkart, J. M., Jaeggi, A. V. & Rudolf von Rohr, C. (2014):** Morality as a Biological Adaptation – An Evolutionary Model Based on the Lifestyle of Human Foragers. In: Christen, M., Van Schaik, C. P., Fischer, J., Huppenbauer, M. & Tanner, C. (eds.): *Empirically Informed Ethics: Morality between Facts and Norms*. Cham: Springer, pp. 65–84.

**Verplaetse, J., Braeckman, J. & De Schrijver, J. (2009):** Introduction. In: Verplaetse, J., De Schrijver, J., Vanneste, S. & Braeckman, J. (eds.): *The Moral Brain: Essays on the Evolutionary and Neuroscientific Aspects of Morality*. Dordrecht: Springer Netherlands, pp. 1–43.

**Voland, E. (2007):** *Die Natur des Menschen: Grundkurs Soziobiologie*. Munich: Beck.

**Voland, E. & Voland, R. (2014):** *Evolution des Gewissens: Strategien zwischen Egoismus und Gehorsam*. Stuttgart: S. Hirzel Verlag.

**Waldmann, M. R., Nagel, J. & Wiegmann, A. (2012):** Moral Judgment. In: Holyoak, K. J. & Morrison, R. G. (eds.): *The Oxford Handbook of Thinking and Reasoning*. Oxford: Oxford University Press, pp. 364–389.

**Wheatley, T. & Haidt, J. (2005):** Hypnotic Disgust Makes Moral Judgments More Severe. In: *Psychological Science*, 16 (10), pp. 780–784.

**Wilson, E. O. (1975):** *Sociobiology: The new synthesis,* 25th anniversary ed. Cambridge, MA: Belknap Press of Harvard University Press, 2000.

**Wilson, T. D. (2002):** *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge, MA: Belknap Press of Harvard University Press.

**Woodward, J. & Allman, J. (2007):** Moral intuition: Its neural substrates and normative significance. In: *Journal of Physiology-Paris*, 101 (4-6), pp. 179–202.

**Young, L. & Saxe, R. (2011):** Moral Universals and Individual Differences. In: *Emotion Review*, 3 (3), pp. 323–324.

**Zimbardo, P. G. & Gerrig, R. J. (2008):** *Psychologie,* 18. ed. Munich: Pearson Studium.

**Zimbardo, P. G., Johnson, R. L. & Weber, A. L. (2006):** *Psychology: Core Concepts,* 5. ed. Boston, MA: Pearson Allyn and Bacon.