Intelligent Automaticity in Moral Judgment and Decision-Making

Asia Ferrin

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2016

Reading Committee:
M. Alison Wylie, Chair
Nancy Snow
William J. Talbott
Sara Goering
Carole Lee

Program Authorized to Offer Degree:
Philosophy

University of Washington

Abstract

Intelligent Automaticity in Moral Judgment and Decision-Making

Asia Ferrin

Chair of the Supervisory Committee:
M. Alison Wylie, Professor
Philosophy and Anthropology

Is conscious reflection necessary for good moral judgment and decision-making? Philosophical attention to this question has increased in the last decade due to recent empirical work in moral psychology. I conclude that conscious reflection is not necessary for good moral judgment and decision-making, arguing that good moral intuition can develop from implicitly acquired and held moral values that bypass deliberative processes.

I anchor my project in the now widely discussed work by Jonathan Haidt (2001) on moral judgment and decision-making. While many theorists think of moral judgment and decision-making as deliberative, Haidt argues that it is in fact more reactive: an individual has a gut response to a moral situation and immediately forms a judgment from that intuition. While Haidt's work is empirically focused, he in large part intends for it to be a challenge to the traditional theory and norms in philosophy. I discuss two kinds of response in the philosophical literature to Haidt's challenge. First, an empirical response: that Haidt's data do *not* show that we largely fail to engage in deliberation. Second, a normative response: that Haidt's empirical data do not challenge our normative ideals. I argue that both groups of arguments fail. This, then, warrants an exploration of the idea that the dominant normative view of moral judgment and decision-making in philosophy needs to be evaluated. I offer my own normative theory of moral judgment and decision-making, which I call "value-guided automaticity," and argue that intuitive judgments motivated by implicitly held values make many of our automatic responses normatively evaluable.

ACKNOWLEDGMENTS

Much thanks is due to the many people who helped my finish this project. First, I would like to thank my family—my parents, Emerald, Dennis, Dakota, Karen, and Mark—for believing in me and encouraging me to pursue my academic career. They have been unwaveringly supportive and positive—in both good times and bad.

I am also grateful for my rock-star dissertation committee. Alison Wylie was one of the first faculty members to get excited about my project and has done an amazing job steering the helm. She has been a strong advocate for my work and career, has taught me much professionally and personally, and possesses fierce integrity and drive that I greatly admire. I have been incredibly fortunate to have her on my team. Alison has also served as a mentor and advisor beyond the dissertation. It was a true honor to work with her for three years at *Hypatia: A Journal of Feminist Philosophy* and she has always provided wise counsel which has helped me thrive professionally, earning positions like Graduate Assistant at the 2012 Philosophy in an Inclusive Key Summer Institute (PIKSI) and Faye Sawyier Predoctoral Teaching Fellow at Illinois Institute of Technology. I am eternally grateful for her guidance and privileged to be one of her students.

Nancy Snow, Bill Talbott, Sara Goering, and Carole Lee have each brought unique expertise and perspectives to my project, for which I am grateful. My project is more diverse and strong because of it. These committee members have also been patient, thorough, and encouraging in approaching my work—I am lucky to have had their challenging, but supportive feedback and attention. They have also been a pleasure to work with as individuals, always brining engaging conversation and laughter to our discussions.

I would also like to thank Ingra Schellenberg for turning me on to my dissertation topic and issues in moral psychology more broadly. I took Ingra's moral psychology course in my first quarter as a graduate student at the University of Washington. She showed me how exciting the topic could be and served as a caring mentor and advocate as I pursued my first two years of coursework at UW.

Other mentors who guided me through the ups and downs of the PhD include Barbara Mack and Karen Emmerman. Both took me under their wings and guided me professionally and personally, determined not only to see me succeed, but thrive during my time at the University of Washington and beyond. I truly would not have made it to this point without their love, wisdom, and compassion. I am grateful not only for their mentorship, but what I know will be life-long friendships.

I would also like to thank the UW philosophy department for their continuous support and the UW Graduate Opportunities and Minority Achievement Program for a generous dissertation fellowship.

Final thanks is due to my undergraduate mentors who encouraged me to pursue philosophy and graduate school—neither of which I had could have imagined as career paths. Special thanks goes to my undergraduate advisor Bridget Newell, Michael Popich, Mary Jo Hinsdale, and the McNair Scholars Program—a program designed to prepare first-generation, low-income college students for graduate studies. I was accepted into the McNair Scholars program the summer after my first year of college and have never looked back. Thanks to the multitude of support from folks over the years, I have achieved my goal and look forward to helping future students do the same.

CHAPTER 1

INTRODUCTION

Shardae is visiting her grandparents, watching a news story about the Afghanistan War. After the story ends, Shardae's grandfather (Steve) goes into the kitchen to get a glass of water. Shardae follows him in and asks whether he thinks the war is morally justified. He pauses to think for a moment, then says no, he does not think the war is morally justified and goes on to give a list of reasons, such as the cost of life and resources, and further loss of security in the US. When Shardae presses: "but why do these things matter?," Steve replies that we should try to act in ways that minimize harm, and these consequences show that the war creates more harm than no war. Steve starts to make himself a sandwich, at which point Shardae goes back into the living room where her grandmother is still watching the TV. Shardae asks her grandmother (Debbie) whether she thinks the war is morally justified. Debbie immediately says no. When Shardae asks her grandmother why she thinks this, Debbie says "It's just wrong. Isn't it obvious?" Shardae presses her grandmother for reasons, but Debbie does not give any, again saying "I don't know, it's just wrong," apparently unable to articulate the reasons for her judgment.

In this scenario, both of Shardae's grandparents make a moral judgment, the same moral judgment, about the war in Afghanistan, namely that the war is unjustified or morally wrong. However, even though Shardae's grandparents make the *same* moral judgment, many are inclined to evaluate the *quality* of their judgments differently. Steve appears to have been more thoughtful and easily provided compelling reasons—based on some moral principles—for his judgment. Debbie, on the other hand, does noes not appear to put any thought into her judgment and seems either unable or unwilling to give reasons for her judgment. Does the process by which they arrived at their judgment make Steve's judgment better?

Is conscious deliberation a necessary feature of good moral judgment and decision-making?[1] This question has received significant attention in philosophy, psychology, neuroscience, and even more mainstream media in the last decade. For example, the John Templeton Foundation recently published "Does Moral Action Depend on Reasoning? Thirteen Views on the Question," which includes contributions by Michael Gazzaniga, Rebecca Newberger Goldstein, Alfred Mele, Christine Korsgaard, Joshua Greene, and Antonio Damasio among others (2010). Much of this conversation about moral reasoning in philosophy has been generated by psychologist Jonathan Haidt's controversial 2001 article, in which he argues that quick, automatic processes largely drive moral judgment while reflective processes play an ad hoc role (Haidt 2001). Haidt characterizes moral judgment and decision-making as reactive: an individual has a gut response to a moral situation and immediately forms a judgment from that intuition. We typically utilize reasoning or reflection, Haidt argues, only when asked to justify our judgments after the fact. In this dissertation, I engage with the current debate about nature and norms of good moral judgment-making.

Haidt argues that moral reasoning and principles *rarely* guide moral judgments and decisions. Haidt takes his argument to be a critique of traditional views and standards of moral judgment-making in both philosophy and psychology. Haidt explains that for many centuries, Western philosophers—and more recently, psychologists—have mostly shared what Haidt calls a "rationalist" conception of good moral judgment-making, which emphasizes the importance of deliberation, impartiality, and appeal to universal truths in human moral psychology. Haidt describes this rationalism:

[1] Note that I use almost interchangeably "judgment," "decision," and "action." While they are obviously different, my and others' analysis often applies to all three. "Judgment-making" and "decision-making" are often talked about as a kind of action.

> Moral psychology has long been dominated by rationalist models of moral judgment…. Rationalist approaches in philosophy stress 'the power of a priori reason to grasp substantial truths about the world' (Williams, 1967, p. 69). Rationalist approaches in moral psychology, by extension, say that moral knowledge and moral judgment are reached primarily by a process of reasoning and reflection (Kohlberg, 1969; Piaget, 1932/1965; Turiel, 1983). Moral emotions such as sympathy may sometimes be inputs to the reasoning process, but moral emotions are not the direct causes of moral judgments. In rationalist models, one briefly becomes a judge, weighing issues of harm, rights, justice, and fairness, before passing judgment on [persons or a moral situation]. (Haidt 2001, 814)

Haidt identifies Plato, the Stoics, medieval Christian philosophers, continental rationalists, Kant, Hare, and Rawls, as endorsing some version of this "rationalistic" view of moral psychology (815-816). Haidt further explains that in the mid-20th century, the dominant paradigm in psychology emphasized the role of emotion in judgment and decision-making, however in the 1970s, Kohlberg—who was largely influenced by Kant[2]—(re)introduced rationalism into psychological theory (as part of the "cognitive revolution" in psychology). Haidt describes Kohlberg's rationalism:

> [Kohlberg] endorsed a rationalist and somewhat Platonic model [of moral judgment and decision-making] in which affect may be taken into account by reason…but in which reasoning ultimately makes the decisions…Kohlberg was quite explicit that the cognitive mechanisms he discussed involved conscious, language-based thinking. He was interested in the phenomenology of moral reasoning, and he described one of the pillars of his approach as the assumption that 'moral reasoning is the conscious process of using ordinary moral language' (Kohlberg, Levine, & Hewer 1983, p. 69). (Haidt 2001, 816)

Haidt continues:

> Kohlberg trained or inspired most of the leading researchers in moral psychology today (see chapters in Kurtines & Gewirtz, 1991; Lapsley, 1996). Rationalism still rules, and there appears to be a consensus that morality lives within the individual mind as a traitlike cognitive attainment, a set of knowledge structures about moral standards that children create for themselves in the course of their everyday reasoning (see Darley, 1993). (Haidt, 816)

[2] Kohlberg writes, for example, "my theory and Rawls's grew out of the same roots; Kant's formal theory in moral philosophy and Piaget's theory in psychology" (Kohlberg 1981, 192; cf Petrovich 1986, 89).

Haidt then proceeds to challenge this "rationalist" view. Before discussing his challenge, let me make two brief notes. First, Haidt would claim that according to this rationalist view Shardae's grandfather's moral judgment is qualitatively better than her grandmother's, as illustrated by Kohlberg's six stages of moral development,[3] for example:

Figure 1

Level & Stage	Age Range	Examples
Preconventional **Stage 1: Avoidance of punishment** **Stage 2: Exchange of favors**	**Preschool – elementary; some junior high; few high school students**	**Stage 1: "I would cheat if I knew I wouldn't get caught."** **Stage 2: "I'll let you copy mine if you do my homework."**
Conventional **Stage 3: Good child** **Stage 4: Law and order**	**Few older elementary children, some junior high, many high school students** **(Stage 4 does not typically appear until high school)**	**Stage 3: "I'm not going to tell because I want her to like me."** **Stage 4: "You can't do that because the teacher said no."**
Postconventional **Stage 5: Social contract** **Stage 6: Universal ethical principle**	**Rarely seen before college (stage 6 is extremely rare)**	**Stage 5: "In this case, the rule may be wrong."** **Stage 6: "You shouldn't lie because it violates the Golden Rule."**

Most relevant is Kohlberg's sixth, and highest, stage of moral development, in which moral judgments and decisions involve a *conscious* appeal to universal ethical *principles*. Shardae's grandfather, Steve, appears to meet the conditions of the sixth stage: he uses his reasoning skills to think about how abstract moral principles (such as the "principle of utility") determine the rightness or wrongness of a particular action (going to war in Afghanistan). Shardae's grandmother, on the other hand, does not even appear to fit into Kohlberg's model.[4] She

[3] Image from "StudyBlue.com" (http://www.studyblue.com/notes/note/n/adol-ch-56-week-4/deck/5395747)
[4] Of course, there have been fantastic critiques of Kohlberg's framework for this reason, such as Carol Gilligan's *In a Different Voice*. Haidt's critique comes at the problem from a different angle, tthus I bracket Gilligan's work here with the intention to explore the issue in future work.

certainly does not engage in any reflection or utilize moral principles. Perhaps she occupies an egoist or conventional rule-based stage. At any rate, she does not appear to excel in moral judgment-making as Steve does.[5]

The second point to note is that Haidt has given a relatively blunt description of rationalism, which may exclude more nuanced rationalistic accounts.[6] He does not go into much more depth than I have presented here and he moves quickly between various philosophical and psychological concepts. However, I am primarily interested in how philosophers have responded to Haidt's work and aim to establish a better theory of moral judgment and decision-making than has currently been offered. Thus, I grant his description and begin with his challenge to this rationalistic view of moral judgment and decision-making (Chapter 1). I then explore and evaluate arguments in moral psychology that respond or relate to Haidt's work (Chapters 2 and 3). Finding both Haidt's model and responses to his work unsatisfactory, I then offer an alternative account of good moral judgment and decision-making (Chapter 4) and explore possible objections to my account (Chapter 5). I turn now, then, to Haidt's challenge to the rationalism described above.

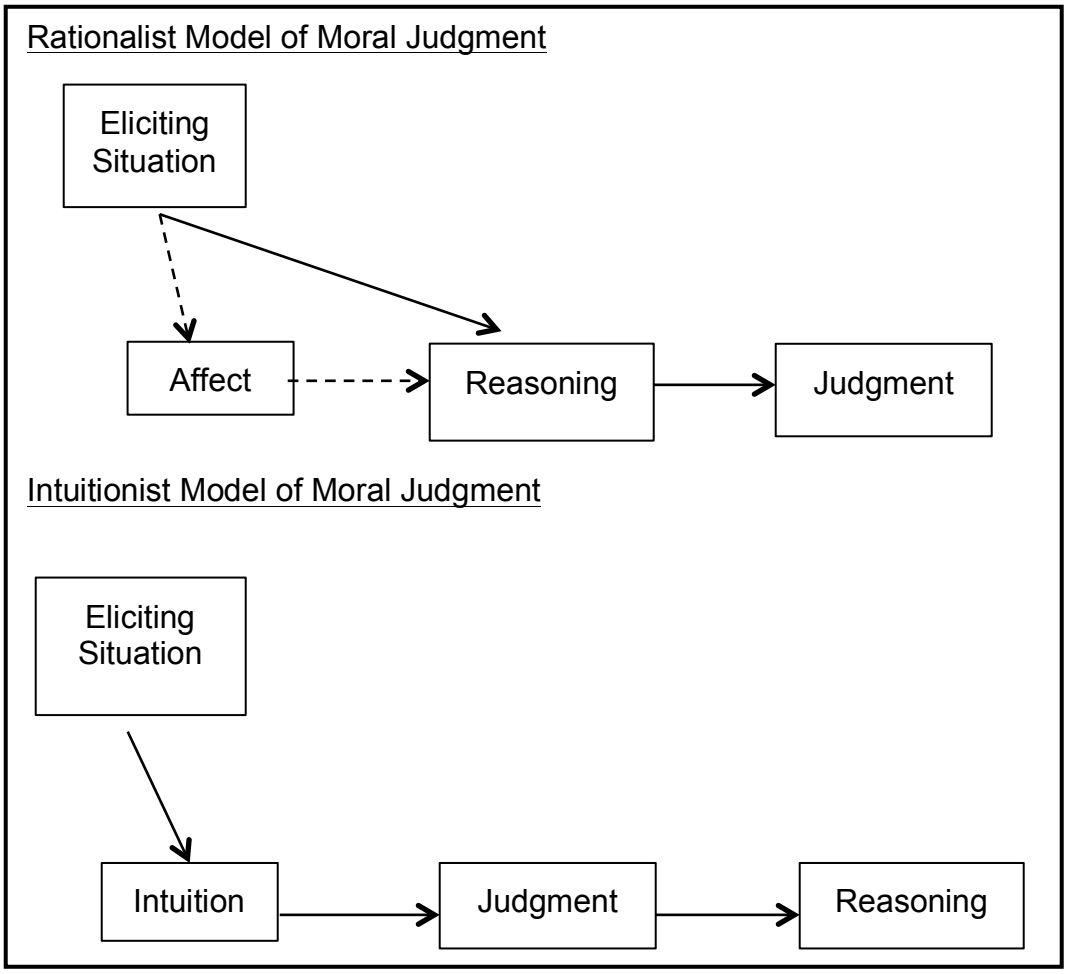
HAIDT'S CHALLENGE TO RATIONALISM IN MORAL PSYCHOLOGY

[5] In most virtue ethics accounts, an agent need not reflect in the moment. If one has so deeply internalized moral reasons over time, she is able to make good moral judgments automatically. Hence, it might be suggested that Debbie does excel in moral-judgment making, according this virtue ethics model. However, notice that according to virtue ethics, even if one does not consciously reflect on moral reasons at the time of judgment, she ought to be able to articulate the reasons if asked (I talk about this point in more detail in Chapter 3); according to virtue ethics, moral reasons directly guide one's judgment and the agent has access to these reasons—presumably, she has spent a significant amount of time reflecting on them in the past. The tension in the scenario about Steve and Debbie is not between someone who consciously deliberates and someone who has internalized moral reasons, but between someone who consciously deliberates about principles and someone who has an unprincipled, automatic gut-reaction.

[6] I discuss this point in more detail in the next chapter.

Haidt makes three primary claims in the paper that has provoked this discussion (2001). First, he argues that our moral judgments and decisions rarely depend upon moral reasoning about abstract principles. Second, he argues that moral reasoning is often post-hoc: individuals instantaneously make a moral judgment or decision and *then* search for reasons to justify their judgment/decision. Third, Haidt argues that moral judgments and decisions are actually driven by an automatic, affect-laden intuitive cognitive system. He calls this a "Social Intuitionist" model of moral judgment. Figure 2 shows this contrast:

Figure 2



Notice that the social intuitionist model of moral judgment inverts the relationship between reasoning and moral judgments.[7] On the *rationalist* model, reasoning leads to a particular judgment; on the *intuitionist* model, a particular moral judgment is the catalyst for moral reasoning. Haidt explains: "The central claim of the social intuitionist model is that moral judgment is caused by quick moral intuitions and is followed (when needed) by slow, ex post facto moral reasoning" (2001, 817). Haidt offers four sources of support for this thesis: "four reasons to doubt the causal importance of reason" (819).

First, Haidt argues that research on cognitive processing suggests that many of our moral judgments and decisions are automatic. Haidt starts with a description of the "dual-process" model of the mind. According to advocates of the dual process model, there are two cognitive systems that guide judgment, decision-making, and action: "System 1" and "System 2." System 1 is the "intuitive system," which is characterized as: fast and effortless, runs automatically, is inaccessible (only results enter awareness), does not demand attentional resources, and is common to all mammals. System 2 is the "reasoning system," which is characterized as: slow and effortful, intentional and controllable, consciously accessible and viewable, demands attentional resources (which are limited), and is unique to humans over age 2 and perhaps some language-trained apes (Haidt 2001, 818; cf. Evans 2008). A quick jump upon hearing a rustle in

[7] Haidt's model is more complex than portrayed in this diagram insofar as there are several social influences/paths as well. He does think that private reasoning can affect intuition ("the reasoned judgment link"), though only in *rare* cases and often when an initial intuition is weak and processing capacities are high (2001, 819). Thus the efficaciousness of private reflection is largely an illusion (819). Change in intuition from private reasoning seems more common when it has an affective component, for example, if I imagine myself trying to protect my children in a war-torn country, I might trigger a new intuition about the ethics of war ("the private reflection link," 819). Haidt also claims that the practice of social reasoning can affect intuitions ("the reasoned persuasion link"): in conversations, one person might give reasons about an issue that trigger a new intuition in the listener (819). Additionally, we can be persuaded through conversation by the motivation to adhere to group norms ("the social persuasion link"). If my neighbor opposes abortion, I might be inclined to do so as well (819). Note, however, that this is a very different understanding of "moral reasoning" than philosophers will have in mind (and different than the System 2 reasoning Haidt describes above), which I discuss in more detail in the next chapter. Furthermore, these additional methods of moral reasoning are not the central focus of his 2001 essay.

the bushes is typically the result of System 1 processing, while arriving at the solution to a difficult math problem would be the result of System 2.

Haidt then explains that many social and cognitive scientists now accept that much of our behavior and judgment is driven by System 1: "The emerging view in social cognition is that *most* of our behaviors and judgments are in fact made automatically (i.e., without intention, effort or awareness of process; Bargh, 1994; Bargh and Chartrand, 1999; Greenwald & Banaji; 1995)" (Haidt 2001, 819; emphasis original). Haidt then discusses some of the findings that support this view. For example, research suggests that people's ultimate judgments about others are largely driven by first impressions. People engage in "thin slicing," meaning they form an immediate evaluation of a person or situation (usually within 5 seconds) based upon relatively limited information. The judgments formed by thin slicing are often the same judgments expressed after longer observation and deliberation. For example, students exposed to an instructor for only 6 seconds are likely to evaluate the instructor *the same* as students exposed to the instructor for an entire semester. Additionally, the judgments formed during thin slicing often influence other judgments and evaluations—for example, the immediate judgment that someone is attractive makes one more likely to judge that person as kind and having good character (Haidt 2001, 820; cf. Thorndike 1920; Dion, Berscheid, & Walster 1972; Albright, Kenny, & Malloy 1988; and Ambady & Rosenthal 1992). Relatedly, Haidt cites the increasingly popular literature on implicit bias, which focuses on thin-slicing based on social stereotypes. For example, many people immediately and implicitly judge a Black face as more dangerous than a white face (Haidt 2001, 820; cf. Devine 1989; Greenwald 1995). Haidt also discusses research on the power and pervasiveness of cognitive-heuristics. He explains:

> According to Chaiken's (1987) heuristic-systematic model of persuasion, people are guided in part by the 'principle of least effort.' Because people have limited cognitive

> resources, and because heuristic processing is easy and adequate for most tasks, heuristic processing (the intuitive process) is generally used unless there is a special need to engage in systematic processing (see also Simon, 1967). (820).

Haidt explains, for example, that people often follow the heuristic: "agree with people whom I like." This heuristic is relevant to our moral judgments because we are more likely to judge someone harshly if a close friend has first expressed harsh judgment toward that person (820). Haidt does grant that in some cases, for example when heuristics or first impressions conflict, conscious reasoning might kick in as we are trying to form judgments (820). However, he claims that most of our day-to-day judgments and decisions, moral and nonmoral, are driven by automatic cognitive phenomena such as thin-slicing, implicit stereotyping, and heuristic-processing.

Haidt's second reason for doubting the causal influence of reasoning on our moral judgments is that much of our reasoning is motivated. In other words, moral reasoning is not a cognitive tool that we use to arrive at truth, but rather a process that serves other social and epistemological ends: "The reasoning process is more like a lawyer defending a client than a judge or scientist seeking truth" (820). Haidt cites various studies to support this claim. For example, experimental results suggest that people show a tendency to search for anecdotes and evidence that *exclusively* support their preferred side of an issue (821; cf. Baron 1995; Perkins et al., 1991). Thus, those convinced of global climate change find themselves reading only about the existence and impacts of increasingly extreme weather patterns, while climate skeptics find themselves reading only about the natural fluctuations in global temperatures over millennia. Haidt further explains that in many situations, once people find a single piece of evidence to support their initial judgment, they stop searching for further information (821; cf. Perking, Allen, & Hafner, 1983). Additionally, people often only give weight to evidence that supports

their judgments. For example, in a study of students who had strong opinions about the death penalty, when students were exposed to evidence on both sides of the issue, they unhesitatingly accepted evidence that supported their view while heavily scrutinizing the evidence that opposed their view. Haidt describes several specific human motives—contrasted to the motive to find truth—that result in this reasoning. First, for example, people are motivated to hold attitudes and beliefs that will satisfy social goals such as harmony and agreement (821; cf. Chen & Chaiken, 1999), and hence are likely to shift their attitudes to align with the supposed views of a potential discussion-partner (821). People are motivated to hold attitudes and beliefs that are consistent, specifically with their self-identity and world-view broadly (821; cf. Moskowtiz, Skurnik, & Galinsky, 1999) and thus search for evidence and reasons to maintain internal-coherence. Such research leads Haidt to conclude that reasoning is often a tool used for justifying one's existing beliefs rather than a tool for arriving at the truth or fact of the matter.

The third reason Haidt gives for doubting that reasoning guides moral judgments is that people often—without awareness or intention—give false or impossible reasons for their judgments, decisions, and behaviors. Haidt describes research where people report reasons for their judgments or behaviors that that could not have actually led to their judgment or decision (822). For example, in one study (Nisbett and Schachter 1966), participants were asked to take an electric shock, where one group was given a placebo pill that was said to produce the same symptoms as electric shock (the other group was not given the pill). Those in the placebo group attributed their heart palpitations and nausea to the pill and were able to withstand four times as much shock as those who had not taken the pill. However, 75% of the participants said they hadn't thought about the pill and made up a variety of reasons for their greater tolerance, for example: "Well, I used to build radios and stuff when I was 13 or 14, and maybe I got used to the

electric shock" (Haidt 822; cf. Nisbett and Wilson, 1977, p. 273). The same kind of thing is shown in studies involving hypnosis, subliminal presentation and split-brain patients. Haidt describes the phenomenon of these "post hoc constructions":

> When asked to explain their behaviors, people engage in an effortful search that may feel like a kind of introspection. However, what people are searching for is not a memory of actual cognitive processes that caused their behavior, because these processes are not accessible to consciousness. Rather, people are searching for plausible theories about why they might have done what they did. (822).

In short, people do not always, perhaps even rarely, have access to the reasons that guide their judgments, decisions, and behaviors and thus (without awareness and unintentionally) "make up" reasons that they think caused their judgment, decision, or behavior. Another study from Nisbett and Wilson 1977 shows that people may create false reasons for decisions. In this particular study, participants were asked to say which pair of nylons—out of a line of four—was the best quality. Many participants picked the pair to the far right. When asked about the reasons for their choice, people cited things like the superior knit, sheerness, or elasticity. However, unbeknownst to the participants, the nylons were all the exact same (Nisbett and Wilson 1977, 243-244; Wilson 2002, 102-103). This study (any many others like it) suggests that people believed they engaged in reasoning to arrive at their choice, but were actually influenced by irrelevant factors (e.g. the position of the nylons). Haidt's point regarding this data is that there is an "illusion of objective reasoning" (822). According to this idea, when Shardae's grandfather, Steve, gives moral reasons and principles as the cause of his judgment, he *thinks* these reasons and principles influenced his judgment, however it is more likely (Haidt suggests) that Steve made a quick judgment and then searched for reasons that would justify or explain that judgment, but did not actually cause the judgment. Steve was under the illusion of objective reasoning, as Haidt would say many of us are.

The final reason for Haidt's skepticism about moral rationalism is that good moral action is more closely associated with *emotional* skills than *reasoning* skills. Haidt discusses research involving psychopaths and people with ventromedial prefrontal cortex (VMPFC) damage to illustrate this point. Psychopaths and VMPFC patients serve as interesting case-studies in moral psychology because they struggle significantly with behaving in moral ways. Psychopaths are commonly thought of as manipulative, egotistical, callous, aggressive, and even maniacal. Some psychopaths engage in extremely harmful behaviors—e.g. killing one's own parents for money—while others are more subtly manipulative and dangerous. Those with VMPFC injuries also—though to a lesser degree than psychopaths—lack concern about social and moral norms, are impulsive, and sometimes manifest extreme and antisocial behaviors (Haidt 2001, 824). Interestingly, psychopaths and VMPFC patients, despite their cognitive damage or atypical development, have species-normal reasoning skills. Haidt explains: "Cleckley characterizes psychopaths as having good intelligence and a lack of delusions or irrational thinking. Psychopaths know the rules of social behavior and they understand the harmful consequences of their actions for others" (824). In other words, psychopaths are capable of engaging in moral reasoning to the same degree as non-psychopaths. Psychopaths can articulate universal moral rules and explain the implications of those moral rules (e.g. killing one's parents violates the harm principle). However, as Haidt explains, "They simply do not care about those consequences" (824). In other words, psychopaths do not lack the capacity for moral *reasoning*, but rather the capacity to feel moral *emotions*. They do not experience remorse, sympathy, shame, embarrassment, love, or grief (Haidt, 824). This lack of emotional engagement—not a failure of moral reasoning—is cited as the cause of morally reprehensible behavior in psychopaths. The same kind of phenomenon is seen in VMPFC patients. Haidt explains:

"Patients with damage restricted to VMPFC show no reduction in their reasoning abilities. They retain full knowledge of moral rules and social conventions, and they show normal abilities to solve logic problems, financial problems, and even hypothetical moral dilemmas" (Haidt 824; cf. Damasio 1994). However, what they lack is emotional responsiveness to morally relevant situations (such as people dying): "When shown pictures that arouse strong skin conductance responses in undamaged people…individuals with VMPFC damage show no response…. These patients know that the images should affect them, but they report feeling nothing" (Haidt, 824). If moral reasoning (System 2) was the primary driver for good moral judgment, we would expect that psychopaths and VMPFC patients would still act in morally appropriate ways because their System 2 functions normally. That these groups fail to act morally, even with their reasoning skills intact, suggests that some other cognitive process (likely housed in System 1) drives good moral judgments and decisions.

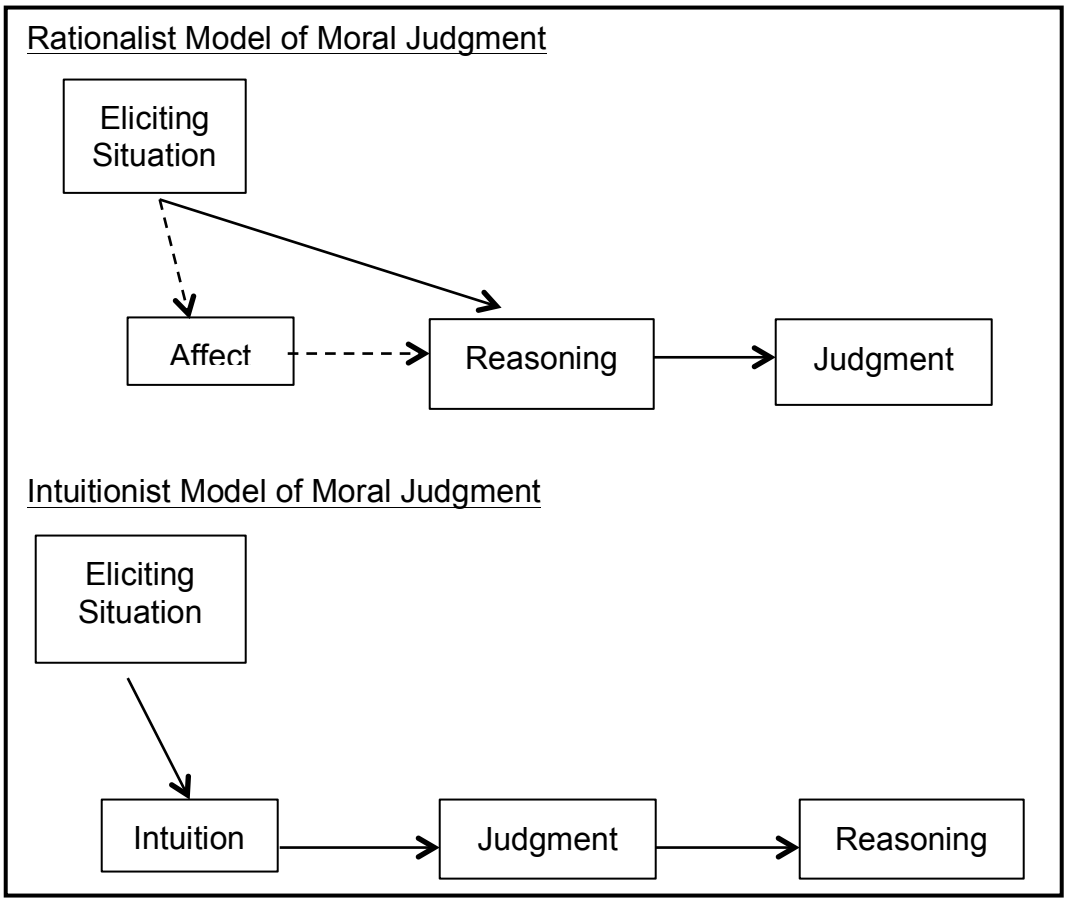
In sum, Haidt argues that the diverse evidence—regarding the influence of System 1 cognitive processing, biases in searching for evidence, post hoc constructions, and the ineffectiveness of reasoning—suggests that moral judgment and decision-making are *not* guided by reasoning and reflection. Haidt claims that the successes of moral judgment and decision-making cannot be understood by further investigation into deliberative processing. He writes: "To really understand how human morality works…it may be advisable to shift attention away from the study of moral reasoning and toward the study of intuitive and emotional processes" (825). Haidt also challenges the idea that moral judgments and decisions are guided by universal, abstract principles. As I will discuss below, Haidt claims that automatic moral judgments are developed and guided by innate and socially constructed moral values. He thus rejects the idea

that moral judgments and decisions are driven by anything like the categorical imperative, harm principle, or any other kind of duty-specific principles.

I now turn to Haidt's alternative theory of moral judgment and decision-making: Social Intuitionism.

HAIDT'S SOCIAL INTUITIONIST VIEW OF MORAL JUDGMENT AND DECISION-MAKING

As discussed above, one of the central claims in Haidt's work is that moral judgments and decisions are largely guided by affect-laden intuitions. Recall **Figure 2** from above:



The figure shows that on Haidt's model, an eliciting situation triggers a moral intuition, which produces a moral judgment, which may then be followed by post-hoc reasoning. I only briefly discuss Haidt's model here to introduce the reader to his account.

In developing this account, Haidt first describes the somatic nature of moral intuitions. They are subtly visceral, perhaps best thought of as "momentary flashes of feeling" (825). For example, when asked about a situation of consensual incest, people have a quick "pang of disgust." This pang of disgust leads them to judge the consensual incest as morally wrong. Haidt explains, though, that these moral intuitions are conditioned and hence over time happen outside of our awareness. Haidt often uses the phrase "gut-feeling" to describe the concept of moral intuition, where a gut-feeling may cause a physiological and behavioral response, but not arise in one's conscious awareness. These gut-feelings may be generated by a web of associations and conceptual schemas. Haidt gives an example:

> …because we all have experience with foods that are easily contaminated, we come to equate purity and cleanliness with goodness in the physical domain. We learn from experience that pure substances are quickly contaminated (e.g., by mold, dust, or insects) when not guarded and that once contaminated, it is often difficult to purify them again. These experiences in the physical world then form the basis (in many cultures) of conceptual schemes about moral purity—for example, children start off in a state of purity and innocence but can be corrupted by a single exposure to sex, violence, drugs, homosexuality, or the devil…. (825)

Haidt concludes: "moral intuition…appears to be the automatic output of an underlying, largely unconscious set of interlinked moral concepts" (825); they are "Momentary flashes of feeling" or "gut-feelings" based upon embodied associations involving moral concepts that lead to an immediate intuition which once made conscious or explicit becomes a moral judgment.

Haidt claims that these moral intuitions are partially innate and partially developed by cultural reinforcement/refinement. Haidt claims that given the social similarities between chimpanzees and humans, there is good reason to believe that humans also possess innate

normative guidelines (826). Haidt further explains that there are four "underlying models of social cognition" which exist across all cultures, providing further evidence for an evolutionarily explanation of the origins of moral intuition: communal sharing, authority ranking, equality matching, and market pricing (Haidt 2001, 826; cf. Fiske 1991, 1992). Additionally, the way these four models develop in children—in threshold-like stages rather than gradually—suggests that they are innate models rather than socially learned (Haidt 2001, 826).

However, moral concepts and values also differ between cultures, which Haidt takes as evidence that our innate moral intuitions are further reinforced or refined by cultural exposure. Haidt claims that there are at least three processes by which cultures "modify, enhance, or suppress the emergence of moral intuitions to create a specific morality" (827). First, a cultural group might emphasize some moral concepts while minimizing others. Haidt cites Shweder's theory of moral ethics which suggests that there are three major clusters, or complexes, of moral concepts: 1) ethic of autonomy (includes values such as rights, freedom of choice, and personal welfare); 2) ethics of community (includes values such as loyalty, duty, honor, respectfulness, modesty, and self-control); and 3) ethics of divinity (includes values such as piety and physical and mental purity) (Haidt 827; cf. Shweder, Much, Mahaptra, & Park 1997). Haidt explains that while an individual is born prepared to develop moral intuitions in all three ethics, typically only one or two is emphasized in her culture. This process parallels that of language acquisition: though our language centers are prepared to develop any language, cultural exposure results in us only developing one to several. The moral complexes are then reinforced, first by cultural immersion in both moral and non-moral contexts and second by peer socialization (Haidt 2001, 827-828).

In sum, Haidt argues that humans have innate moral intuitions that are further shaped through cultural processes. These moral intuitions are surprisingly embodied. When confronted with a moral situation, humans experience an affect-laden gut feeling which triggers the moral intuition, which produces a moral judgment or decision. Crucially, moral reasoning (as typically understood) does not often play a role in this process. The moral intuitions are innate, they are implicitly reinforced/refined, and they are triggered without our awareness. Haidt believes that such a model of moral judgment and decision-making best coheres with empirical research in the social sciences on moral judgment, decision-making, and actions.

Haidt does not take his article to be the final word on moral judgment and decision-making. Instead, it is a call for further research. Haidt writes, for example:

> This review is not intended to imply that people are stupid or irrational. It is intended to demonstrate that the roots of human intelligence, rationality, and ethical sophistication should not be sought in our ability to search for and evaluate evidence in an open and unbiased way. Rather than following the ancient Greeks in worshipping reason, we should instead look for the roots of human intelligence, rationality, and virtue in what the mind does best: perception, intuition, and other mental operations that are quick, effortless, and generally quite accurate (Gigerenzer & Goldstein, 1996; Margolis, 1987). (Haidt 822)

Haidt makes three specific suggestions for further exploration of social intuitionism: research on the quality of moral judgments when reasoning is directly interfered with; research in more diverse contexts and formats, diverging from interview-based studies; and more research across the disciplines (829-830). He concludes by saying "The time may be right…to take another look at Hume's…thesis: that moral emotions and intuitions drive moral reasoning, just as surely as a dog wags its tail" (830).

Moving Forward

While Haidt's work is empirically focused, he in large part intends for it to be a challenge to the traditional theory and norms in philosophy. Both contemporary and past philosophers have placed high value on deliberation and principles. Here, I have outlined three kinds of responses to Haidt's challenge to rationalism and defense of social intuitionism:

1. **Reject Haidt's Claim as False**. On this response, one would argue that Haidt's challenge fails because it is descriptively or empirically false. Haidt uses empirical data to show that we do not make moral judgments according to a rationalist ideal. This justifies, on his view, a rejection of moral rationalism and the consideration of a new theory of moral judgments, grounded in social and cultural norms. One could argue, however, that Haidt's data does *not* show that we fail to make moral judgments according to the rationalist ideal. This might be because Haidt has failed to consider counter-evidence, or has considered only one of several possible explanations of the data, or has over-generalized from his case studies to moral judgment and decision-making more broadly. If Haidt's descriptive challenge fails, he gives no reason to reconsider philosophical descriptive or normative theory.
2. **Reject the Normative Significance of Haidt's Analysis**. Here, one would argue that Haidt's empirical data do not matter normatively. In other words: normative ideals do not need to be altered by the empirical data Haidt presents. As long as it is still *possible* for us to manifest the rationalist ideal, we should continue to strive for that ideal. Haidt has only shown that it is very difficult to reach this standard, which is so much the worse for us.
3. **Embrace Automaticity**. Here, one would take up Haidt's call to further explore the nature and philosophical implications of automaticity. Given the evidence that we operate very differently from the ideal established by moral rationalism, we can use this as an opportunity to think about the value of the standard in the first place. In doing so, one might argue that

the rationalist standard of good moral judgment-making is problematic, and hence ought to be replaced with a new normative standard.

In this dissertation, I explore and evaluate each of these categories of responses. In Chapter Two, I consider various arguments for the claim that Haidt's challenge fails at the *empirical* level. In Chapter Three, I review arguments for the claim that Haidt's challenge does not necessitate a need to revise our *normative* ideals. I argue that both types of response fail. This, then, warrants an exploration of the idea that the normative theory in philosophy needs to be revised. In Chapter Four, I offer my own normative theory of moral judgment and decision-making that takes into account the considerations raised in Chapters 2 and 3. In the final chapter, I discuss the most pressing potential objections to my account.

Works cited

Bargh, J. A. 1994. The Four Horsemen of automaticity: Awareness, efficiency, intention, and control in social cognition. In R. S. Wyer, Jr., & T. K. Srull (Eds.), Handbook of social cognition (2nd ed., pp. 1-40). Hillsdale, NJ: Erlbaum.

Cohen, Rachel. 2004/2010. "Hume's Moral Philosophy." Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/entries/hume-moral/

Evans, Jonathan St. B. T. 2008. Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*. 59: 6.1-6.24

Gazzaniga, Michael. 2010. Does action depend on moral reasoning? John Templeton Foundation. *The Big Question Series*.

Goldstein, Rebecca. 2010. Does action depend on moral reasoning? John Templeton Foundation. *The Big Question Series*.

Haidt, Jonathan. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108 (4).

Kohlberg, Lawrence, and William P. Alston. 1970. *From is to ought: How to commit the naturalistic fallacy and get away with it in the study of moral development*.

Korsgaard, Christine. 1996. *The sources of normativity*. Cambridge; New York: Cambridge University Press.

Korsgaard, Christine. 2010. Does action depend on moral reasoning? John Templeton Foundation. *The Big Question Series*.

Nisbett, Richard E., and Timothy D. Wilson. 1977. "Telling more than we can know: Verbal reports on mental processes". Psychological Review. 84 (3): 231-259.

Peters, R. S., and C. A. Mace. 1967. Psychology. In *The encyclopedia of philosophy.*, ed. Paul Edwards. Vol. 7. New York: Macmillan.

Piaget, Jean. 1965. *The moral judgment of the child*. New York: Free Press.

Wilson, Timothy D. 2002. *Strangers to ourselves: discovering the adaptive unconscious.* Cambridge, Mass: Belknap Press of Harvard University Press.

CHAPTER 2

Introduction

In Chapter One, I summarized Jonathan Haidt's argument that deliberation and moral principles rarely guide moral judgment and decision-making and that automatic, affective-laden intuitions actually drive most of our moral judgments and decisions. In philosophy, there has been significant skepticism about Haidt's thesis. In Section I of this chapter, I explain and analyze such skepticism. I discuss three specific worries about, or objections to, Haidt's analysis. First, one might argue, our daily experiences indicate that—contrary to Haidt's claim—we *do* engage in internal deliberation often about moral issues. Second, our daily experience indicates that we engage in genuine deliberation *with others*. And finally, one might argue that Haidt's analysis only shows that much of our deliberation and moral commitments are automatized, which is consistent with many popular and accepted accounts of moral psychology in philosophy. While these points are compelling, I argue that they nevertheless miss the point of Haidt's analysis. This is due, I argue, in part to biases about the importance of deliberation and in part to Haidt. In the second section of this chapter, then, I explain in more detail why philosophers have missed Haidt's point. In the third section of this chapter, I attempt to reconcile this tension by pulling out and reframing the promising element of Haidt's work. This project is best captured by Haidt's claim that:

> This review is not intended to imply that people are stupid or irrational. It is intended to demonstrate that the roots of human intelligence, rationality, and ethical sophistication should not be sought in our ability to search for and evaluate evidence in an open and unbiased way. Rather than following the ancient Greeks in worshipping reason, we should instead look for the roots of human intelligence, rationality, and virtue in what the mind does best: perception, intuitions, and other mental operations that are quick, effortless, and generally quite accurate (Haidt 2001, 822)

I review the current state of work on automaticity in psychology and cognitive science in effort to evaluate the potential impact of automatic processing on philosophical theories of moral judgment and decision-making. I conclude that current work on automaticity shows that philosophers ought to take seriously Haidt's suggestion—if not his own analysis—that we should be looking for the roots of human intelligence, rationality, and virtue in automatic processing.

Section I: Skepticism About Haidt's Account

A number of philosophers have written in response to Haidt's work, arguing that it is wrong or flawed.[8] In this section, I discuss three major reasons for such skepticism. Recall that Haidt claims that "... moral judgment is caused by quick moral intuitions and is followed (when needed) by slow, ex post facto moral reasoning" (817). The first response to Haidt's claim is that there are in fact, as indicated by our experience, many cases where moral judgments or decisions are caused not by quick moral intuitions, but by moral reasoning. For example, imagine that I have plans with a friend for this weekend, but I am suddenly feeling less inclined to go. I am very tired and had a stressful encounter with a co-worker, so I do not feel up for a social activity. However, I may reason that I should not cancel with my friend because it would leave her hanging and may be disappointing as she was looking forward to seeing a particular movie with me. I do think in general it is morally good to follow through on commitments, as doing so indicates a respect of others' time and the relationship. I then realize that I have a tendency to flake out with this friend and thus I am perhaps especially obligated to follow through on our plans this weekend if I can. There is even a chance, I think, that it might be restorative to go out and put aside the long week for a bit. I spend a few moments thinking about all of these

[8] See, for example, Pizarro and Bloom 2003, Saltzstein and Kasachkoff 2004, Fine 2006, Horgan and Timmons 2007, Clarke 2008, Musschenga 2008, Kennett and Fine 2009, Craigie 2011, Kennett 2012, and Sauer 2012, for example.

considerations: I am tired, but it might make me feel better; I don't feel like going, but also do not want to disappoint my friend. I really should be better about keeping my commitments… After some mulling, I finally decide that I will go to the movie in effort to be a good and reliable friend, but if I am not feeling well, might ask that we cut the evening short.

This looks like a genuine case of deliberation that leads to my moral decision or action (going to the movies may not seem moral, but keeping my commitment is). Darcia Narvaez, writing in response to Haidt, lists some of her own experiences with such everyday deliberation (as it interacts with automatic and affective responses): "He looks upset; what could it be; what should I say?" "I suppose I should stop over there and say hi, but I don't feel like it." "Did I handle the kids well enough? What would be better next time?" "I can't believe I am expected to use my time this way. How can the system be changed?" (2008, 235). With these examples, Narvaez aims to show that deliberation about moral issues occurs often in our everyday experiences. Of course, this deliberation may be subtle and somewhat spontaneous—happening while walking to the office, during your commute, or as your mind wanders away from the text you are reading. But other times it may be central in your focus: "Okay, I need to decide what to do about my co-worker's inappropriate comments" or "I need to figure out how to better handle David disrupting my class."[9]

[9] See, for example, Pizarro and Bloom 2003 on such a point: "When one looks outside the laboratory…there is considerable evidence that people do struggle with moral issues, trying to determine the right thing to do. Coles (1986), for instance, documented the moral struggles faced by Black and White children in the American South during the Civil Rights movement, recounting the decisions they made; Gilligan (1982) did similar work looking at young women who are deciding whether to get an abortion…many moral issues are personal and have to be addressed by each individual in the course of his or her life: How much should I give to charity? What is the proper balance of work and family? What are my obligations to my friends? There are no 'off-the shelf' answers to these questions, no immediate gut reactions as to what is right and wrong. Haidt (2001) was likely correct that we do have quick and automatic responses to certain situations—killing babies, sex with chickens, and so on—and although these responses can be modified and overridden by conscious deliberation, they need not be. But most moral cognition is not about such simple cases; in the real world, moral reasoning is essential" (195).

We also appear to engage in deliberation about more high stakes moral issues. A friend, for example, was recently trying to decide whether to get a "first trimester screen," a less invasive procedure that pregnant women over 35 can now select instead of amniocentesis. She took a number of considerations into her decision—what were the risks to the fetus, how might she benefit from the testing, what would she do with such information, what are the social ramifications of such testing—and thought about the issue for weeks. There are a number of other major moral life events that persons might deliberate about: which political representatives and referendums to vote for, where to send one's children to school, how to best care for one's aging parents, whether to stay in a particular relation/partnership, whether to forgive those who have wronged us, etc. These examples suggest—again, contra Haidt—that not only do we engage in genuine deliberation about various day to day moral matters, but also about larger life-changing moral events.[10]

A second objection to Haidt's work is that not only do we engage in this internal private reflection about various low- and high-stakes issues, but also in deliberation with others. In fact, the main methodology of academic philosophy is, arguably, a continuous dialogue with others in pursuit of better understanding, answers, or solutions to problems. We deliberate with others in courses, at colloquia and conferences, and through published writing. Furthermore, outside of philosophy, we often deliberate with others about various problems or questions. Take my example above about whether to cancel my plans for this weekend with my friend. If I am feeling particularly unsure about what to do, I might ask my partner about it over dinner—"I

[10] Jeanette Kennett (2012) makes such a point: "Whether to end or stay in a marriage, whether to give a child diagnosed with attention deficit/hyperactivity disorder (ADHD) Ritalin, whether to put an elderly parent into a nursing home or care for him or her oneself, whether to accept a job offer in another city and move children away from their school and their friends—these are decisions with significant moral dimensions, and they are the kind of decisions most of us will face. They are not usually snap decisions. We tend to spend a lot of time thinking about them and canvassing the options before deciding what to do, and we often engage in a process of reevaluation and revision after they are made. The past is a country we often return to, and on each visit we find something new" (259).

can't figure out whether I should cancel or keep my plans; can you help me?"—or I might call other people I trust to get additional information about how my friend might respond or to see what they would do if in my situation. In short, I deliberate about my decision with others. Also consider the above example of my friend thinking about a first trimester screen. Part of her process in arriving at that decision was to talk with me, not just as a friend, but also as someone who had taught and thought about such issues. In short, it seems clear that—contra Haidt—we *do engage in deliberation*, not only through private reflection, but also through genuine dialogue with others.[11]

Haidt does claim that conversation with others is a significant part of moral judgment and decision-making. He explains that there are six major "links" between judgments, intuition, and reasoning in his social intuitionist model. One of the links is called the "reasoned persuasion link" (Haidt 2001, 818). Haidt explains that we do sometimes engage in dialogue with others about reasons, and sometimes such dialogue results in a change of judgment. The idea, then, is that we can affect others' judgments by offering reasons—which would seem consistent with the claim that we deliberate with others. However, his idea here is much different than the kind of dialogue I described above. It is presumed in my description above that one changes her judgment, or arrives at a particular decision, because she is rationally responding to the reasons raised by her interlocutor. When I ask my partner about whether I should cancel my weekend plans, I am asking for reasons that would be in favor of one choice over the other. But, Haidt does not think this is what actually happens in conversation; we are not moved to new judgments or decisions by the force of reasons. Instead, Haidt explains, the giving of reasons may trigger

[11] Saltzstein and Kasachkoff 2004 make such a point: "We often experience a change in our moral intuitions regarding a particular situation when someone points out features of the situation that we had not noticed or persuades us to appreciate the importance of some feature that we had not given sufficient consideration. And persuading us to notice something or to appreciate its significance is often achieved by means of argument" (276).

new or different intuitions, which would then lead to new judgments through subconscious automatic processing. I am not, then, responding to the actual reasons given by my partner, but rather having a new intuition triggered. Haidt explains: "Because moral positions always have an affective component to them, it is hypothesized [by the social intuitionist model] that reasoned persuasion works not by providing logically compelling arguments but by triggering new affectively valences intuitions in the listener" (819). Hence, this "reasoned persuasion" on Haidt's view is actually quite different than what many might presume. It *is* a challenge to his view, then, to protest that actual reasons-responsive deliberation with others occurs in our everyday moral lives.

The third major challenge to Haidt's account may grant the prevalence of automaticity in our moral judgment and decision-making, but point out that much of those automatic intuitions are developed by explicit deliberation. Hence, Haidt's sweeping claim that automaticity guides much of our moral judgment and decision-making, while deliberation is primarily post-hoc, is overstated and does not challenge the existing philosophical literature. It is well-accepted within much of the contemporary and historical philosophical work on the topic that it would be impossible for deliberation to guide each and every one of our moral judgments and decisions. Instead, we internalize moral commitments, values, or guidelines that allow us respond automatically in the moment. This internalization happens through explicit deliberation. For example, I might think about what exactly it means to be a good sister. Perhaps my sister and I recently had a conflict and I want to be more supportive in the future. I ask myself: how exactly do I achieve this? I deliberate about the question over an extended period of time and arrive at some particular judgments: I should be more patient, I should listen more attentively, I should support my sister's decisions even when I disagree with them, etc. Deliberation leads me to these

conclusions. I will then need to practice acting according to my new commitments and will struggle to get it right at first. I might have to keep these commitments present in my conscious awareness in order to properly act upon them. But after some time, I will have internalized these commitments (previously produced by my deliberation) and will act upon them automatically—without deliberation—from this point forward. I will not have to explicitly think about how to respond each time my sister needs help; I will respond automatically and often without conscious awareness, affirming her decision, for example, without even realizing it. On such a view, Haidt is correct that much of our moral judgment and decision-making is automatic, but that does not mean it is not guided by deliberation.[12] To the contrary, much of our automatic responses depend upon, or are developed by, explicit deliberation.[13]

Section II: Stalemate in the Conversation

Unfortunately, these challenges to Haidt's work miss the mark and as a result the conversation reaches somewhat of a stalemate. Haidt has several responses readily available for the above challenges. First, Haidt would suggest that we only *think* we engage in deliberation about moral issues. Haidt's precise point in his work is that, contrary to popular belief and our perceived experiences, data suggest that we are often fooled into thinking we have deliberated when in fact we have merely reacted or been influenced by factors outside of our awareness. Hence, to argue

[12] See, for example, Hanno Sauer 2012 on this point: "I will argue that the automaticity of moral judgment can be squared with its rationality just in case moral judgments are based on patterns of moral reasoning that have, through a process of moral education, become habitual. I will show why it is legitimate to think of habits—acquired automatic processes—as processes that can be placed in the space of reasons, and I will make good on the claim that moral judgments are in fact based on such educated intuitions" (259). And a related, but slightly different point from Pizarro and Bloom 2003: "Prior reasoning can determine the sorts of output that emerge from these intuitive systems. This can happen through shifts in cognitive appraisal, as well as through conscious decisions as to what situations to expose oneself to. In both of these regards, prior controlled processes partially determine which fast, unconscious, and automatic intuitions emerge. " (194).

[13] A related idea is that moral judgments and decisions are often automatically processed, but engage with implicitly held moral principles or commitments (not necessarily developed through deliberation). See, for example, Terry Horgan and Mark Timmons 2007 and Tania Lombrozo 2009. I do not engage these arguments directly since the authors are not committed to the idea that deliberation does play a significant role in moral judgment and decision-making.

that "our experiences tell us that we do in fact deliberate," only confirms how strong our self-deception about deliberation is. If Haidt's point is that we engage in deliberation (moral reasoning specifically) much less often than appears to the naked eye, claiming that I/we often deliberate proves to be an unconvincing response to Haidt's thesis.

Haidt's response here, however, moves the conversation in an unproductive direction. The conversation meets a standstill; one side insisting "but I do deliberate," and the other resisting "no, you only think you do." In these terms, there is no obvious way to resolve the disagreement. I suspect that this standstill occurs because both sides fail to take seriously the claim of the other: defenders seem to fail to take Haidt's evidence seriously—as the evidence *does* suggest that our experiences are often misunderstood—and Haidt fails to take seriously genuine experiences of deliberation. I will talk about this problem in more detail below.

A second response that Haidt might offer is to grant that deliberation does happen as I've described in some of the cases above, but it is incredibly rare. He would say in fact, that this kind of moral deliberation mostly occurs among philosophers.[14] Hence, while philosophers might respond that deliberation *is* a common practice, for most people in most places at most times, such deliberation does not occur. Haidt writes:

> People may at times reason their way to a judgment by sheer force of logic, overriding their initial intuition. In such cases reasoning truly is causal and cannot said to be a 'slave to the passions.' However, such reasoning is hypothesized to be rare, occurring primarily in cases in which the initial intuition is weak and processing capacity is high. (2001, 819)

[14] "However, people are capable of engaging in private moral reasoning, and many people can point to times in their lives when they changed their minds on a moral issue just from mulling the matter over themselves. Although some of these cases may be illusions…other cases may be real, *particularly among philosophers, one of the few groups that has been found to reason well* (Kuhn, 1991)" (2001, 819; emphasis mine). Haidt claims here that philosophers' experiences are the exception. One might take Haidt's point here to show that people *can* engage in deliberation with the proper training, though this does not necessarily follow from Haidt's point (it could be that those naturally inclined to engage in reflection end up in philosophy) and is also a separate point than the one Haidt is grappling with. Haidt is theorizing about how people (in general/as a whole) *do* make moral judgments and decisions, not how they *could* or *should*.

Haidt's point is that deliberation is so rare and gut-reactions so common, that research programs in moral psychology ought really to be focused on how automatic processing guides moral judgment and decision-making. Haidt doesn't think that philosophers are necessarily wrong when pointing to examples of deliberation, but that philosophers are pointing to the exception rather than the rule.

However, the conversation here also reaches somewhat of a standstill. Haidt claims that moral deliberation is rare, philosophers reply that moral deliberation is more common than Haidt allows, Haidt responds that philosophers are confused and deliberation is actually rare. This disagreement makes central the question of which occurs more, deliberation or automatic reaction?[15] This question is not only nearly impossible to answer (meticulous tallying would be needed) but also somewhat boring.[16] It doesn't matter for many research programs whether deliberation occurs in only 20% of our moral judgment and decision-making or in 80%. At this point in moral psychology, theorists are simply trying to get a better understanding of the cognitive processes that underlie judgment and decision-making. Both deliberation and

[15] Haidt frames the conversation in this way in response to Darcia Narvaez's 2010 commentary on Haidt's work: "This brings us to a second way of framing the debate: as a contest between models of the partnership between reasoning and intuition. There are three such models: 1. Reasoning as a senior partner. Intuition and emotion are acknowledged, but most of the action is in moral reasoning, which can 'channel' moral emotions, and which can and ought to drive moral behavior. This was Kohlberg's view. 2. Equal partnership. Both processes are (roughly) equally important in our daily lives, and both can work independently to reach different conclusions. This is Narvaez's view…. 3. Intuition as a senior partner. Reasoning is acknowledged, but most of the action is in moral intuition, which can 'motivate' reasoning, and which often drives moral behavior. This is my position, which was shaped strongly by the work of David Hume, Robert Zajonc, Antonio Damasio, John Bargh, and Richard Shweder" (2010, 183). Haidt then goes on to argue, unsuccessfully in my view, for #3.

[16] The question is interesting only if meant to have normative implication. The motivation of one's argument that we *do in fact* often deliberate might be to show that our status is preserved as competent moral judgment and decision-makers. But this latter claim—that we are competent moral judgment and decision-makers—is a normative point. It entails the normative assumption that one must engage in deliberation in order to produce *morally worthy* moral judgments and decisions. I challenge this assumption in the next chapter. I suspect, as I discuss more below, that this normative assumption in large part drives many of the descriptive concerns about Haidt's work. Hence, if the assumption was warranted, this question of whether reasoning or intuition is the primary force of moral judgments and decisions might be more interesting. However, as I argue, the assumption is not warranted. Furthermore, none of the authors objecting to Haidt's work seem aware of nor make explicit the assumption, so many of the arguments thus far presented against Haidt's work do result in a simple disagreement over which occurs *more*, deliberation or automaticity?

automaticity play a large enough role that both warrant attention. Hence, this "deliberation occurs a lot"/ "no deliberation is rare" debate detracts from progress. It may further be argued that given how little we currently know about automatic processing, research on automaticity is important to prioritize and invest in, but this is a much more modest claim that need not depend upon the claim that deliberation rarely, if ever, occurs. In a way, then, Haidt and his interlocutors are pushing a point that need not be pushed and is counterproductive insofar as it divides rather than unites researchers working on the topic.

Finally, in response to the suggestion that automaticity is prevalent, but is built up by deliberation and practice, Haidt will again respond that philosophers only think this, but in fact our automatic processes are developed by evolutionary and social forces.

In sum, the above objections to Haidt's work do not successfully defeat his view. He anticipates and addresses these claims in his work, the thrust of his argument being that deliberation is often an rare and an illusion. One may press that the above objections are not intended to defeat Haidt's view, but to illustrate that the role of deliberation ought to be taken seriously. However, Haidt would reply here that his intention is not to defeat deliberation, but rather to illustrate that the role of automatic processing ought to be taken more seriously. Why have philosophers resisted this call?

First, I suspect that philosophers have been slow to embrace research on automatic processing because psychologists have been slow to produce such research. Psychologist Timothy Wilson (2002) explains that there has historically been surprising resistance to accept the existence of unconscious processing and pursue further research on the topic in academic psychology. He explains that many of the recent findings about automatic processing were discussed by several major figures in the *mid 19th century*, but largely ignored for at least 100

years. Wilson shows that theorists such as William Hamilton (philosopher), Thomas Laycock (neurophysiologist), and William Carpenter (physiologist) were writing about the unconscious in the 1860s, arriving at conclusions that could be mistaken for the claims in modern day psychology journals, for example: lower-order mental processes occur outside of awareness, much thought happens by automatic processing, and we harbor unconscious prejudices which can be stronger and more pernicious than conscious prejudices (Wilson 2002, 10-11). Only in the last several decades have psychologists been seriously interested in these ideas raised over 150 years ago.[17] Wilson hypothesizes that a delay of investigation into the unconscious happened in psychology because late 20th century psychologists did not want to associate themselves with Freud and his theory of the unconscious. Furthermore, as behaviorism developed in response to Freud's study of the unconscious, psychological study focused more on behaviors/outputs rather than the workings of the mind (13). Wilson concludes, then: "Without this backdrop, it is possible that psychology would have discovered sooner than it did that the mind, including the nonconscious mind, can be studied scientifically" (13). In short, research about unconscious (System 1) processing in psychology has been largely avoided until very recently due to biases in academic psychology.

I further suspect that particular assumptions and biases in philosophy have exacerbated this existing hesitancy. It is often presumed that one does not need any tools or special skills to practice philosophy; one must simply be able to think.[18] As is becoming increasingly obvious, even young children can practice philosophy—they can reason about what it means to be a good

[17] See also, for example, Bargh and Chartrand (1999): "What was noted by E.J. Langer (1978) remains true today: that much of contemporary psychological research is based on the assumption that people are consciously and systematically processing incoming information in order to construe and interpret their world and to plan an engage in courses of action" (462).

[18] Of course, people might practice philosophy better or worse depending on their skills and experience, but the point is that *anyone* can practice philosophy as long as they can think about reasons.

friend, what it means to be fair, how we come to know the world, etc. In fact, it is often thought that other activities and cognitive functions outside of deliberation interfere with the practice of philosophy—to do good philosophy, I need to quell distractions, quiet my mind, and sit with a book or my word processor. I then call to mind different reasons and evaluate those reasons.[19] Sometimes, I might engage in this activity with other people, though that activity serves the same purpose: calling to mind reasons and evaluating them. The research on automaticity, however, suggests that, first, we do not actually have the ability to call to mind many reasons or ideas. Many reasons operate outside of our conscious awareness and cannot be easily accessed, if at all. Second, the research suggests that even if we can call to mind reasons, we cannot evaluate them without bias. Such unbiased evaluation is difficult not simply because of our values or philosophical orientation ("As a trained Kantian, it is difficult for me to see those positive outcomes as reason in favor of the new policy"), but because of biases in our motivations, desires, preferences, and emotional states. The research on automatic processing suggests we only see the tip of the iceberg when we engage in deliberation—a host of other cognitive processes continuously operate outside of our awareness, constantly influencing our conscious calling to mind and evaluation of reasons. Even if we could access those other influences, it would be very hard—given the sheer complexity—to pull them apart from our deliberation. In a way, then, Haidt's work can be read as a challenge to philosophical practice. People who have devoted their lives and/or careers to reflection of reasons might thus be skeptical of the argument

[19] Some philosophers would argue that this is a mistaken picture of philosophical practice. Philosophy is not an unbiased investigation of arguments, but rather an exchange of ideas based in each of our experiences. I take it that this is a less popular, though by no means incorrect, view of philosophy however.

that it is very difficult, in some cases impossible, to genuinely reflect on ours and others' reasons.[20]

Not only does work like Haidt's challenge the norms of philosophical practice, but also a deeply held philosophical idea: that "reason" is superior to "emotion/intuition." This idea is prevalent in historical philosophical texts. Of course, many contemporary theorists have arrived at more tempered views about the nature of the relationship between reasoning and affective states—affective states foster better reasoning; reasoning and emotion work in tandem in moral judgment and decision-making; good moral judgment and decision-making cannot occur without affective states; etc.—however, our ability to reason is often given primacy. I talk more about the idea that reasoning *must* be the primary driver of our moral judgment and decision-making in the next chapter. Suffice it to say here that I take this philosophical idea to create an unwarranted resistance to Haidt's work. I suspect that even for those who may find the stronger claim that reasoning is superior to emotion/intuition unpalatable, it still functions for many as an implicitly held belief such that, even if one endorses more intuition-friendly views, one cannot help but make judgments and decisions that are influenced by the reason/emotion hierarchy (just as even die-hard egalitarians often cannot help but sometimes be influenced by oppressive biases). Again, I will say more on the topic in the next chapter.

It is not only philosophers who have reason to maintain skepticism about the prevalence of automatic processing in our lives. The idea that our judgments and actions might be guided by mental processes outside of our conscious awareness does not square well with how humans understand themselves, and hence, we may be motivated to reject a thesis like Haidt's. A common reaction among my students (mostly non-philosophy majors), for example, when first

[20] See also, for example, John Doris (2015): "A preoccupation with reflection is, arguably, the Western philosophical tradition's most distinctive feature, in both historical and contemporary contexts" (17-18).

introduced to Haidt's view is that his work is offensive. However, Haidt's work is not "offensive" as in rude or disrespectful. But, I do think my students find his work insulting; that is, it is insulting to them that Haidt suggests they are reactive, intuitive, unreflective beings. This may explain why some students have such a strong emotional, not just intellectual, response to Haidt's work. I had one student who said he had never been so offended or felt so strongly about an idea in his life. This was shocking to me given that I teach highly controversial topics in my other courses. However, I have noticed that the topics that students find most offensive in my other courses are those that threaten their personal identities: for example, essays that are critical of white privilege, masculinity, heteronormativity, etc. Thus, while it is surprising to see students react in the same way to Haidt's work as to Marilyn Frye's, the connection may be that both threaten their personal identities, i.e. the kind of people they take themselves to be.

Timothy Wilson offers a similar hypothesis about why people might resist acknowledging the influence unconscious processing has on our judgments and decisions. Wilson makes the following two claims in Chapter 5 of *Strangers to Ourselves*: "Many human judgments, emotions, thoughts, and behaviors are produced by the adaptive unconscious" and "Because people do not have access to the adaptive unconscious, their conscious selves confabulate reasons for why they respond the way they did…" (2002, 106). Wilson suggests that there may be resistance to accept such claims because to acknowledge them implies that we are less in control of our judgments and decisions that we currently believe. He writes: "One explanation is that it is important for people to feel that they are the well-informed captains of their own ship and know why they are doing what they are. Recognizing [these above claims] is likely to make people feel less in control of their lives, a feeling that has been shown to be associated with depression" (2002, 113). I think it is possible that Haidt's work creates a kind of

anxiety for people: "Haidt cannot be right because then the control I have over my judgments and decisions is largely an illusion." One may find oneself on the edge of an existential crisis.

However, these kinds of biases are not solely responsible for a resistance to Haidt's work. Haidt has not made it easy to accept his work, for three reasons. First, Haidt presents his work in an adversarial tone. I have suggested that people often seem bothered by Haidt's work, and I suspect that this is in part his intention. In a response to Saltzstein and Kasachkoff's critiques of his work, for example, Haidt cheekily writes things like: "Although I am not originally from the South, I do feel the need to rise up and object to this affront to my honor!" (2004, 283), "I now respond to S&K's challenge by firing three shots. Each shot is aimed at one of the three fundamental errors I see in S&K's portrayal of the [social intuitionist model]. Readers who are not interested in the duel may skip ahead to the refreshments afterward" (284), and "That's better…I have defended the honor of the social intuitionist model. I can now return to a more genteel and hospitable mood in which I can entertain some of the ideas raised by S&K" (286). Of course, Haidt is being playful here, but I nevertheless think it is significant that he encases his theoretical work in this adversarial tone because while the tone is explicit here, I think it is implicit in his work as well. For example, in his 2001 essay (his most popular and most referenced work), he offers a *brief* history of rationalism in philosophy and psychology. Not only is this summary underdeveloped (which could be frustrating to any philosopher) he presents it under the hyperbolic heading of "Philosophy and the Worship of Reason" (2001, 815) in contrast to "Psychology and the Focus on Reasoning" (816). Though this is subtle and small evidence, I think it indicates that Haidt aims to alienate philosophers in his work.[21]

[21] Of course, I could be wrong here, in which case I warmly welcome a correction that clarifies how Haidt's work is intended to produce a constructive dialogue rather than a two-sided war.

Furthermore, Haidt unfairly represents the view he aims to challenge: "rationalism." He writes for example: "Rationalism—the view that reason is the chief source of valid knowledge—has long been based on faith and hope, not on observations of actual human behavior" (2010, 183). Clearly, Haidt has overstated his point. In the first section of this chapter, I described a number of cases where one might infer—*from one's own observations and experiences*—that reasoning plays a significant role in our moral lives. Recall how I deliberated about whether to cancel my plans and my friend struggled with the choice of prenatal testing. Our experiences and observations of human behavior suggest that reason *does* play a large role in our knowledge- and decision-making. Haidt's point, unfortunately, is not only inaccurate, but again divisive—those who identify as rationalists might understandably take offense to Haidt's suggestion here that rationalist views lack rigor or any sort of empirical grounding, based only instead of "faith and hope." Haidt also fails to show appreciation for the plurality and complexity of rationalist views. In his description of rationalism in philosophy, he cites only Plato, the Stoics, Medieval Christian philosophers, the 17th century's continental rationalists ("e.g., Leibniz, Descartes"), Kant, R.M. Hare, and Rawls (Haidt 2001, 815-816). These theorists, however, represent only a small fraction of theorists who might be called rationalists. Furthermore, even if Haidt were to qualify that he intends only to address these specific rationalist views, he fails in all of his work to give any sort of in-depth discussion about the details of the above theorists' accounts. Hence, he presents somewhat of a caricature of rationalism and then targets that caricature. Understandably, then, philosophers may hesitate to give Haidt's analysis the consideration it might otherwise deserve.

Finally, Haidt misrepresents the potential of automatic processing. He presents automaticity as atavistic rather than intelligent, dynamic, and morally responsive. As I explained in the previous chapter, Haidt makes the following claims about the process of moral judgment

and decision-making: we most often have affectively-laden gut-reactions, from which our moral judgments and decisions directly follow; [22] those intuitions are primarily developed and influenced by social interactions,[23] norms[24] and values;[25] when we engage in reasoning, it is often only to justify our intuitively-developed judgments and decisions;[26] and intuitions can change over time due to social influence, but rarely (if ever) in response to moral reasons. In other words, our socially developed intuitions are automatically and subconsciously triggered to create conscious moral judgments and decisions that we rationalize after the fact. On Haidt's description, our moral judgment and decision-making does appear to be primarily reactive and insensitive to moral reasons. However, Haidt claims that this is a mischaracterization of his view. This issue comes up in Haidt's reply to Saltzstein and Kasachkoff's critique. In his reply, Haidt claims that Saltzstein and Kasachkoff have oversimplified his model:

> All three of these misreadings make sense once you realize that S&K are critiquing a stripped-down version of the SIM that I label here the "possum" model (Figure 1). In this model, evolution built a bunch of intuitions into people's heads, and when people are confronted with social situations, these intuitions fire off, causing judgments, which cause post hoc reasoning. In this two-link model, reasoning plays no causal role in the judgment process, we are all prisoners of our gut feelings, and it is hard to see how societies advance or individuals change their minds. (Haidt 2004, 283)

[22] "The model proposes that moral judgments appear in consciousness automatically and effortlessly as the result of moral intuitions" (Haidt 2001, 818).
[23] "[Verbal] reasoning can sometimes affect other people, although moral discussions and arguments are notorious for their rarity with which persuasion takes place. Because moral positions always have an affective component to them, it is hypothesized that that reasoned persuasion works not by providing logically compelling arguments but by triggering affectively valenced intuitions in the listener" (Haidt 2001, 819).
[24] "Because people are highly attuned to the emergence of group norms, the model proposes that the mere fact that friends, allies, and acquaintances have made a moral judgment exerts a direct influence on others, even if no reasoned persuasion is used" (Haidt 2001, 819).
[25] "[The model] proposes that morality, like language, is a major evolutionary adaptation for an intensely social species, built into multiple regions of the brain and body, that is better described as emergent than as leaned yet that requires input and shaping from a particular culture. Moral intuitions are therefor both innate and enculturated" (Haidt 2001, 826).
[26] "The model proposes that moral reasoning is an effortful process, engaged in after a moral judgments is made, in which a person searches for arguments that will support an already-made judgment" (Haidt 2001, 818).

Haidt goes on to claim that this is far from the model that he presents. However, this *is* in large part the model that he presents. He objects however, claiming, for example, that reasoning *does* play a large role in his model (contra Saltzstein and Kasachkoff):

> I do indeed minimize the causal efficacy of private reasoning, in which a person thinks about an issue and questions assumptions, beliefs, and intuitions without the benefit of a discourse partner. I allow for this possibility in Link 5, but I cite evidence (e.g. Kuhn, 1991; Perkins, Farady, & Bushey, 1991) that, in general, people other than philosophers are bad at such reasoning. Ordinary people do not spontaneously look for evidence on both sides of a judgment question. (2004, 284)

First, this is precisely the point that Saltzstein and Kasachkoff contend with. Nevertheless, Haidt continues:

> But moral reasoning *does* play an important causal role once it is seen as a social activity rather than as a solitary activity. People engage in moral reasoning not so much to figure things out for themselves, in private, but to influence others. (2004, 284)

Here, Haidt uses the term "reasoning," but clearly means something very different than what Saltzstein and Kasachkoff, most philosophers, and most non-philosophers mean by *reasoning*, as I noted above.[27] Giving reasons to try to persuade others does not typically count as "moral reasoning." Thus, Haidt's claim that Saltzstein and Kasachkoff have misrepresented his work is unfair. Haidt equates "trying to convince others" with "evaluating the merits of reasons." But simply calling them both "reasoning" clearly does not make them the same.[28] Hence, I maintain that my above description of Haidt's work is accurate.[29] And I further suggest that his

[27] Haidt equivocates again: "Yet ever since Plato wrote his Dialogues, philosophers have recognized that moral reasoning naturally occurs in a social setting, between people who can challenge each other's arguments and *trigger new intuitions*…" (Haidt 2001, 820; cf. Haidt 2004, 285; emphasis mine). Haidt is mistaken that philosophers think philosophical dialogue is about triggering new intuitions and hence mistaken that his view is consistent with the traditional connotation of "moral reasoning in social settings."

[28] Peter Railton (2014) suggests that Haidt's view has evolved over the last decade (813, footnote). I think this is mistaken as Haidt's later work (2012, 2013 for example) confirms his original model presented in "The Emotional Dog and Its Rational Tail" (2001).

[29] Neil Levy (2006) offers the same interpretation: "Haidt shares with the emotivists the view that moral argument works, when it works, but persuasion, not reason: by bringing others to share one's emotions, rather than by the logical force of reasons" (100). See also Joseph Paxton and Joshua Greene (2010): "…according to [Haidt's Social Intuitionist Model], social influence on moral judgment only occurs when one person succeeds in modifying another's intuition. In other words, the SIM includes no social counterpart to the "reasoned judgment" link, which

oversimplification contributes to the strong resistance to his view. Haidt claims that his "review is not intended to imply that people are stupid or irrational" (Haidt 2001, 822).[30] However, I do not think Haidt achieves this intention. Haidt describes a socially responsive, but not morally responsive, automatic cognitive system. However, the empirical data suggest that System 1 processes are much more responsive and dynamic than Haidt describes. In the next section, then, I aim to correct for Haidt's mistake here by showing the intelligence and usefulness of automatic processing in our moral lives.

I have spent a significant amount of time in this section attempting to explain various players' motivations and biases because I think this is useful for understanding the nature and future of the conversation about automatic processing in moral judgment and decision-making. The current set-up of this dialogue has resulted in somewhat of a stalemate. My aim is to reframe the set-up of the conversation so that we may push through this stalemate to achieve a widespread and productive research program on the nature and norms of moral judgment and decision-making. As I mentioned in my introduction I do aim, in a sense, to build upon Haidt's work. He makes an important call for theorists working on moral judgment and decision-making: "Rather than following the ancient Greeks in worshipping reason, we should instead look for the roots of human intelligence, rationality, and virtue in what the mind does best: perception, intuitions, and other mental operations that are quick, effortless, and generally quite accurate" (Haidt 2001, 822). While I think the framing and details of Haidt's work are problematic, I

---

would allow one persons' moral reasoning to influence another 's moral judgment directly, without first modifying the target's intuition…Indeed, depending on what one means by 'reasoning,' one might say that the SIM does not really allow for 'reasoned persuasion' at all" (513-514)

[30] It is also not obvious that his is intention is not to imply that people are stupid or irrational. He cites numerous studies that show we are responsive almost entirely to morally irrelevant reasons and are largely motivated and biased in our judgment and decision making (Haidt 2001). He even writes in later work: "My claim is that each of us is flawed as an individual reasoner. We are each cursed by the 'confirmation bias' (the tendency to seek only evidence that will confirm our pre-existing beliefs [Nickerson, 1998] and nobody has yet figured out a way to 'debias' people (Lilienfeld, Ammirati, & Landfield, 2009)" (Haidt 2013, 288). Haidt does go on to claim that better reasoning can be achieved in social situations, but his view of us as individual reasoners is nevertheless dispiriting.

nevertheless take seriously his call for revised perspectives and research in moral psychology. Those who have objected to Haidt's work have been able to easily dismiss his call for a revision of this status quo. This, I think, is a mistake.

Section III: The Current Status of Automaticity in Moral Judgment and Decision-Making

Research on the influence of automatic processes has boomed since Haidt's 2001 essay. We know many things about moral judgment and decision-making that were previously unthinkable, for example that smell, taste, body language and position, and even color influence our moral judgments and decisions. That said, there is still much we do not know about the relationship between automaticity and moral judgments and decisions. In this section, I offer an overview of the current empirical research on automaticity and deliberation.

Jonathan Haidt's work on automaticity and moral judgment and decision-making starts with a dual-process theory of mind. According to this theory, discussed above, there are two main types of processes that operate in the brain to guide all of our physical and mental activity. Haidt explains:

> It must be stressed that the contrast of intuition and reasoning is not the contrast of emotion and cognition. Intuition, reasoning, and the appraisals contained in emotions (Frijda, 1986; Lazarus, 1991) are all forms of cognition. Rather the words *intuition* and *reasoning* are intended to capture the contrast made by dozens of philosophers and psychologists between two kinds of cognition. The most important distinctions (see Table 1) are that intuition occurs quickly, effortlessly, and automatically, such that the outcome but not the process is accessible to consciousness, whereas reasoning occurs more slowly, requires some effort, and involves at least some steps that are accessible to consciousness. (Haidt 2001, 818).

Here is the "Table 1" Haidt refers to in this passage:

Features of the intuitive system	*Features of the reasoning system*
Fast and effortless	Slow and effortful
Process is unintentional and runs automatically	Process is intentional and controllable
Process is inaccessible; only results enter awareness	Process is consciously accessible and viewable
Does not demand attentional resources	Demands attentions resources, which are limited
Parallel distributed processing	Serial processing
Pattern matching; thought is metaphorical, holistic	Symbol manipulation; thought is truth preserving, analytical
Common to all mammals	Unique to humans over age 2 and perhaps some language-trained apes
Context dependent	Context independent
Platform dependent (depends on the brain and body that houses it)	Platform independent (the process can be transported to any rule following organism or machine)

Haidt's work depends upon this distinction; again, he argues that the intuitive system largely drives moral judgment and decision-making, while the reasoning system plays a secondary role by either triggering the intuitive system or rationalizing the judgments and decision produced by the intuitive system. It is worth getting clear on, then, current views about this distinction and its implications.

Psychologists Bertram Gawronski, Jeffrey Sherman, and Yaacov Trope provide an overview and status update of dual-process theories in their introduction to their edited anthology *Dual-Process Theories of the Social Mind* (2014). In this review, they explain the function of dual-process theories, key features of dual-process theories, and differences and similarities amongst dual-process theories. First, Gawronski et al. claim that the aim of dual-process theories is to explain the underlying mechanisms of observed cognitive and psychological input-output relations. For example, as I am watching a scary movie, my heart rate increases, my brow furrows, I hold my movie-watching companion more tightly, and in certain moments I cover eyes with my hands. Dual-process theories seek to explain the mental mechanism(s) that cause

these outputs to result from the relevant input (the scary movie). Dual-process theories specifically hold that there are two distinct (sets) of mental processes underlying my varying input-output relations: "A central feature of dual-process theories is that they postulate two qualitatively distinct (sets of) mental processes that mediate between inputs and outputs" (4). This is in contrast to single-process and multiple-process theories. While there are advocates of single- and multiple-process theories, Gawronski et al. argue that dual-process theories prove most useful and best explain empirical evidence, suggesting that single-process theories are too general and abstract to be helpful (13), while multiple-process theories are too specific.[31] Gawronski et al. conclude: "For many social psychological questions, the conceptual distinctions proposed by dual-process theories have clearly demonstrated their value in explaining and predicting the phenomena of interest" qualifying that "However, for other questions, more fine-grained theories may be needed to account fully for the available evidence" (14).

Gawronski et al. describe three different kinds of dual-process theories: dual-process theories, dual-representation theories, and dual-system theories (7). Dual-process theories "emphasize functionally distinct mental processes" (7), characterized as an automatic mental process and nonautomatic mental process (3).[32] For example, the initial categorization of a person is automatically processed, while more detailed integration of individual features is nonautomatically processed (8). The automatic and nonautomatic processes are characterized as

[31] Gawronski et al. reject the idea that empirical evidence could confirm the number of mental processes that exist, focusing instead on the usefulness of the various frameworks in explaining empirical data: "Yet when discussing the question of how many processes there 'really' are, it is important to note that existence claims—including claims about the existence of one, two, or multiple processes—are ontological in nature. In the philosophy of science, ontological claims fall in to the realm of metaphysics, which means they cannot be tested empirically (e.g., Popper, 1934; Quine 1960). From this perspective, it is not possible to test whether there are one, two, or multiple processes.

[32] Actually, many theorists contrast "automatic" with "controlled" processes. However, Gawronski, Sherman, and Trope find this distinction inaccurate: "Automatic processes are often contrasted with controlled processes. Yet the term *control* has been used to refer to either (1) a particular feature of nonautomatic processing (i.e., controllability) or (2) an umbrella concept subsuming multiple different features of nonautomatic processing (i.e., awareness, intentionality, resource dependence, controllability). To avoid conceptual confusion, we use the term *nonautomatic* as the semantic antonym of the term *automatic* instead of the more common term *controlled* (see Moors & De Houwer, 2006)" (16, endnote 1; emphasis original).

unconscious/conscious, unintentional/intentional, efficient/resource-dependent, and uncontrollable/controllable, respectively (6).[33] In contrast, *dual-representation theories* do not emphasize distinct mental *processes*, but rather distinct mental *representations*. For example, an Implicit Association Test might show that while I have explicitly egalitarian views about women, I nevertheless exhibit negative stereotypes and associations implicitly. According to a dual-representation theory, this dissonance happens because I have two different representations of the subject (women), one is manifested explicitly, the other implicitly (8). Finally, Gawronski et al. describe a third category of dual-process theory: *dual-system theory*. Dual-systems theory is the kind of dual-process theory Haidt describes above. Gawronski et al. explain: "The shared assumption of such *dual-system theories* is that multiple psychological dualities are systematically correlated, thereby constituting two functionally distinct mental systems" (8; italics original). Gawronski et al. then list a number of psychological dualities that closely match Haidt's table included above:

> Although dual-system theories differ in their assumptions about which dualities represent core features of the proposed systems, the hypothesized correlations between dichotomous characteristics are often depicted in lists of features that describe one of the two systems as *associative, automatic, slow learning, experiential, affective, parallel,* and *holistic,* and the other one as *rule-based, nonautomatic, fast-learning, rational, cognitive, sequential,* and *analytic*…. (8; italics original)

Again, these dichotomies closely resemble those presented by Haidt. Gawronski et al. cite a number of social psychologists who endorse this kind of dual-process theory, including Seymour Epstein (1994), Steven Sloman (1996), Elliot Smith & Jamie DeCoster (2000), Daniel

[33] There is disagreement about the extent to which the four different features of automatic/nonautomatic processing co-occur and how this affects dual-process theory (Gawronski et al. 2014, 5). While some theorists argue that the presence on one feature is enough to characterize a process as either automatic or nonautomatic, Gawronski et al. correctly identify difficulties with such a view (e.g. that a process could be simultaneously automatic and nonautomatic) (6). On the other hand, some theorists suggest that because the four features do not necessarily co-occur, the categories should be broken down to the 16 different possible permutations of the four features, challenging a dual-process theory altogether (6). I do not see that Gawronski et al. satisfyingly address this issue.

Kahneman (2003), Fritz Strack & Roland Deutsch (2004), and Robert Rydell & Allen McConnell (2006).

In *Rationality and the Reflective Mind* (2011), psychologist Keith Stanovich also indicates that there has been wide-reception over the past few decades of a dual-systems model like the one Haidt describes. He explains that while there is much variety among dual-system theories, there is nevertheless growing consensus that the mind can be characterized by two different types of cognition, which he says have different functions and different strengths and weaknesses (16). He presents the following Table to show the extent, diversity, and similarity of such dual-process theories:

Keith Stanovich, *Rationality and the Reflective Mind*, Table 1-1 (pg 18):

TABLE 1.1 Some Alternative Terms for Type 1 and Type 2 Processing Used by Various Theorists

Theorist	Type 1	Type 2
Bargh & Chartrand (1999)	automatic processing	conscious processing
Bazerman, Tenbrunsel, & Wade-Benzoni, (1998)	want self	should self
Bickerton (1995)	online thinking	offline thinking
Brainerd & Reyna (2001)	gist processing	analytic processing
Chaiken et al. (1989)	heuristic processing	systematic processing
Evans (1984, 1989)	heuristic processing	analytic processing
Evans & Over (1996)	tacit thought processes	explicit thought processes
Evans & Wason (1976; Wason & Evans, 1975)	type 1 processes	type 2 processes
Fodor (1983)	modular processes	central processes
Gawronski & Bodenhausen (2006)	associative processes	propositional processes
Haidt (2001)	intuitive system	reasoning system
Johnson-Laird (1983)	implicit inferences	explicit inferences
Kahneman & Frederick (2002, 2005)	intuition	reasoning
Lieberman (2003)	reflexive system	reflective system
Loewenstein (1996)	visceral factors	tastes
Metcalfe & Mischel (1999)	hot system	cool system
Norman & Shallice (1986)	contention scheduling	supervisory attentional system
Pollock (1991)	quick & inflexible modules	intellection
Posner & Snyder (1975)	automatic activation	conscious processing
Reber (1993)	implicit cognition	explicit learning
Shiffrin & Schneider (1977)	automatic processing	controlled processing
Sloman (1996)	associative system	rule-based system
Smith & DeCoster (2000)	associative processing	rule-based processing
Strack & Deutsch (2004)	impulsive system	reflective system
Thaler & Shefrin (1981)	doer	planner
Toates (2006)	stimulus-bound	higher order
Wilson (2002)	adaptive unconscious	conscious

Stanovich claims that while there is great diversity in this table, there is also "family resemblance" (18), namely that Type 1 processes are autonomous, meaning they automatically fire when triggered by an external stimuli and do not depend on higher-level input (19), while Type 2 processes are nonautonomous, meaning they can be intentionally activated or deactivated and do depend on higher-level input (20).

Clearly, there is plenty more that can be said here, but my aim is simply to show that current resources like Gawronski, Sherman, and Trope's *Dual-Process Theories of the Social Mind* and Stanovich's *Rationality and the Reflective Mind* suggest that the distinction Haidt draws upon between System 1 and System 2 continues to be widely accepted in social and cognitive sciences.

Furthermore, there is a growing consensus on the degree of influence each of these processes on our judgments and decisions. In one of the essays from *Dual-Process Theories of the Social Mind*, Roy Baumeister and John Bargh—psychologists who have written extensively on automaticity—aim to summarize the current evidence for and status of conscious and unconscious processing. They describe their project:

> We, the authors of this chapter, have found ourselves on opposite sides of debates about several important questions, included the efficacy of conscious thought and the scientific viability of free will. Still, we have followed each other's work over the years with interest, respect, and admiration, and this has enabled our programs of research to benefit and to be informed by each other's work. Moreover, we actually agree on far more than our periodic debates might suggest. Our purpose in this chapter is to explore and elucidate these areas of agreement. (2014, 35).

Baumeister and Bargh first describe a "spectrum" of views on the relationship of conscious and unconscious processing. On one end of the spectrum are "complete control" views. According to such views, our judgments, decisions, and behaviors are under control of our conscious processes—we have access to our reasons, motivations, internal and external influences, goals,

etc. Baumeister and Bargh claim that while this view still has much intuitive appeal, few serious researchers endorse it given the overwhelming evidence of unconscious processes and influences (36). On the other end of the spectrum is the view that consciousness is completely irrelevant to behavior. On this view, our judgments, decisions, and behaviors are entirely driven by unconscious automatic processes. Consciousness is epiphenomenal—it does not play an influential role on our mental states and behaviors, instead functioning as a side effect of our unconscious processing (36). Baumeister and Bargh explain that this view has had a number of proponents in the last decade including Daniel Wegner (2002), Timothy Wilson (2002), and Ap Dijksterhuis and Loran Nordgren (2006). [34]

In between these two extremes, Baumeister and Bargh describe two major views that regard conscious and unconscious processes as more complementary rather than competing. On the first view, conscious thoughts are highly influential in guiding judgment and behavior, but unconscious processes can shape the content of consciousness. For example, a commercial for gardening might remind you of your brother and you may consciously decide to give him a call to catch up. While the external stimuli automatically triggered an association for you, you were able to consciously notice the association and decide what to do about it. According to Baumeister and Bargh: "This position is amenable to the commonsense view that conscious thoughts are ultimately in charge of action, but it assigns an important role to unconscious processes as providing support and input" (36). A slightly different view—endorsed by Baumeister and Bargh—is that unconscious processes do most of the work to guide our behaviors, while conscious processing occasionally enters the picture to alter the unconscious stream of behavior (36-37). The idea here is that consciousness plays more of a supporting rather

[34] Haidt would be included here.

than lead role in human behavior. Baumeister and Bargh introduce a car metaphor to help elucidate these competing views:

> The full conscious control metaphor would suggest that consciousness is the car's driver, who works the controls so as to direct the car toward his or her intended destination. The [consciousness as largely irrelevant] view would depict consciousness as a passenger, perhaps in the back seat. The passenger may have a rich subjective experience of the journey, but is simply seeing what happens, without having any influence on where the car goes. [On Baumeister and Bargh's view] consciousness is akin to a fancy navigational system. Unconscious processes mostly drive the car, but occasionally they do not know how to get where they want to go, so they consult the navigational system, which can perform calculations that the driver cannot… (37).

In short, Baumeister and Bargh claim that both automatic and conscious processes play important and different roles in guiding judgments, decisions, and behaviors. Specifically, Baumeister and Bargh suggest that unconscious processes guide much of our in-the-moment decisions and actions, freeing up our conscious processes to think about past and/or future experiences (43).

Another example of an important role that conscious processing plays is correcting inaccurate or problematic automatic judgments or actions. For example, one of the more successful strategies one can use for combating implicit bias is to create and practice implementation intentions (Gollwitzer 1999). Implementation intentions are "If-then" statements that help counteract implicit stereotypes or associations one may have. Implementation intentions are consciously developed and practiced. For example, if I have recently learned that I have negative stereotypes of women—perhaps I associate women with emotion instead of logic—I might create an implementation intention to combat this negative stereotype. My implementation intention could be "If I encounter a woman at the workplace, I will give her comments and ideas full consideration" or "If I there is a political conflict with a female colleague at work, I will give her the benefit of the doubt until evidence proves otherwise." The success of such

implementation intentions illustrates an important role for conscious processes in our judgment and decision-making.[35]

Such results convincingly show that reasoning can intervene on automatic processes, to a greater degree than Haidt himself allowed. Again, Haidt describes automatic processes as largely resistant to intervention, or nonresponsive to genuine reasoning. The research described above suggests he may have overestimated or overstated this fact. Hence, a more moderate claim is best supported by the evidence: *automatic processes in some contexts are less responsive to System 2 reflection, but in many cases can be influenced—either shaped or overridden—by conscious deliberation.*

While this research points to an important role for conscious processing, it is worth noting that unconscious processing can also be used to successfully correct problematic implicit biases. For example, studies show that negative implicit biases can be combated not only with explicit deliberation, but also implicit exposure to positive counter-stereotypes. For example, simply being in a room with counter-stereotypical images (a poster of a Black university president or a picture of a female scientists or philosophers) can counter the influence of negative implicit biases and associations (Blair 2002). Thus, while conscious deliberation can be a tool for intervention, it is not the only tool.

This latter point is important because it helps undermine the popular idea that conscious processing serves primarily to regulate our, in large part, bad automatic judgments and decisions: "those automatic processes are too blunt and thus often get us into trouble; conscious processing

[35] Interestingly, conscious overriding of problematic implicit biases seems to only be successful when in the form of these implementation intentions. Direct conscious interference with implicit biases—e.g. "No, women are not emotional, that is an incorrect association I have—has proven less successful in controlled studies (Gollwitzer 1999). This fact suggests that conscious deliberation is not all-powerful and, even when able to influence judgments and decisions, is still restricted by the boundaries and influence of automatic processes.

keeps automaticity in line and takes care of the finer-detailed/higher-concept cognitive work."[36] I take this to be an inaccurate and unnecessary framing of automatic and conscious processes. While unconscious processes do sometimes result in judgments or actions that we would not consciously endorse—an anti-racist still finds himself crossing the street to avoid a Black man at night—unconscious processes also serve us very well. For example, unconscious processes can give us an accurate "sense" about a situation before our conscious processes are able to diagnose the situation. This phenomenon has been demonstrated through the Iowa Gambling Task (IGT) (Bechara et al. 1994, Turnball et al. 2014). In the IGT, participants are instructed to draw cards from four decks. The cards in each of the decks give either a monetary "reward" or "punishment"—for example, one card might be +$120, while another card would be -$80. The cards in each deck appear random to the participant, though each of the decks has a different overall value. Deck 1, for example, might have higher overall losses while Deck 3 has higher overall reward. Participants are instructed to draw from any of the decks, but to do so in a way that maximizes their profit. Interestingly, many participants get an implicit "sense" for which decks are overall good and which are overall bad very quickly. This "sense" is indicated by a) physiological reactions to the decks—for example, increased sympathetic nervous system activity when participants hovered over the wrong decks and b) the fact that participants could

[36] See, for example Moore and Loewstein (2004): "This paper argues that self-interest and concern for others influence behavior through different cognitive systems. Self-interest is automatic, viscerally compelling, and often unconscious. Understanding one's ethical and professional obligations to others, in contrast, often involves a more thoughtful process. The automatic nature of self-interest gives it a primal power to influence judgment and make it difficult for people to understand its influence on their judgment, let alone eradicate its influence. This dual-process view offers new insights into how conflict of interest operate and it suggests some new avenues for addressing them or limiting some of their greatest dangers" (189); Gendler 2008: "In the final section of the paper, I turn to the topic of how we might regulate and respond to discordant alief in cases where discord is unwelcome. As beings who are simultaneously embodied and capable of rational agency, the challenge is one that we face repeatedly. This has not gone unnoticed. It is the challenge that the ancients explored when they considered the problem of *harmonizing the parts of the soul*, and that the moderns discussed when they examined the *conflict between reason and the passions*. And it is one that contemporary cognitive and social psychology (among other disciplines) have been exploring under many rubrics—both behavioral and neurological" (Gendler 2008, 554). Gendler goes on to discusses "two strategies for regulating [automatic processes]" (554), by which she means bringing them into accordance with our reflective attitudes.

report affective responses to each of the decks after only 20 cards and correctly identify the good and bad decks after 50 cards. Participants, however, cannot explain the pattern of the decks or exactly why some decks are better than other. They are only able to report their affective responses and general evaluations of each of the decks. Researchers take these results to indicate that our implicit, affective processes help us make good judgments and decisions.[37] This kind of research gives less credence to the idea that the role of conscious processing is to regulate pesky automatic processes. To the contrary, automatic processes can be responsive and give us accurate information to help produce good judgments and decisions.[38]

Relatedly, conscious processes sometimes produce undesirable judgments and behaviors, and need to be regulated by the unconscious system (inverting the above assumption about the superiority of conscious processing). For example, image that in a presidential election you find, much to your surprise, that you like the primary candidate of the party you typically oppose. You cannot really allow yourself to accept this, though; while you feel an affinity for the candidate, you keep telling yourself consciously that you would never vote for a Republican. However, it turns out that your affinity for the candidate is due to the fact that she seems to share many of your same core values and commitments. Hence, you "have a good feeling about her," and for good reason. However, your conscious commitments to a particular political party inhibit you from appreciating the similarity and result in your rejecting her as a viable candidate, even

[37] The IGT specifically compares cognitively normal participants against participants who have VMPFC damage. Those with VMPFC damage are unable to get an accurate sense of the decks and thus select more from worse than better decks.

[38] See also, for example, Ambady 2010 on *thin slicing*. In one study, participants were asked to rate a professor's effectiveness after watching six seconds of a silent video of him/her. The study found that the participants' ratings based on this "thin slice" correlated to the *actual students' end-of-the-semester ratings* of the professor. In a similar study, several instructors offered a lesson and participants were tested afterwards on the material. Other participants then evaluated each instructor's effectiveness based on a few seconds of a silent video. In this study, the participants' ratings of instructor effectiveness correlated with *actual instructor effectiveness*, measured by previous participants' test scores. In these studies, participants are able to pretty accurately evaluate someone after only a few seconds of exposure. Research suggests this "thin slicing" skill extends to many other contexts and domains and is entirely based on System 1 processing.

though her taking office would be in line with your values and interests. In such a case, conscious processing has led you astray while unconscious processes have worked successfully. Take another example: imagine that you have had a conflict with a good friend. Perhaps he betrayed you by telling a secret you he had explicitly and sincerely promised to keep. When you consciously think about the situation, you tell yourself that you cannot forgive him. His action has caused a significant problem for you and you believe that you cannot trust him given the breach. However, you find that whenever you see him in the neighborhood or at social events, you miss your friendship and feel compelled to forgive him. But you quickly activate your conscious processes to remind yourself that he made a bad choice that cannot be undone and thus should not (at least at this point) forgive him. However, your heartstrings tug as you walk away from him at the dinner party. One reason your heartstrings might tug is because you actually have good reason to forgive him, but those reasons are not present in your conscious thinking at the moment. Your friendship runs deep, he seems genuinely sorry, and you actually do not want to have distance in the relationship. But again, you might not have conscious access to these considerations, and thus through your deliberation you arrive at the conclusion that you should not forgive your friend, even though your automatic processes rightfully nudge you in that direction.[39] In sum, while conscious deliberation can sometimes correct problematic judgments and decisions driven by automatic processes, automatic processes can also sometimes correct problematic judgments and decisions arrived at via conscious deliberation.[40] It is not the case that the primary role of conscious processes is to override or correct those hopelessly erroneous automatic judgments and decisions.

[39] See Karremans and Aarts 2007 for more on how automaticity promotes forgiveness of close others.
[40] I discuss these kinds of cases and the point here in more detail in the next chapter.

In fact, recent empirical evidence suggests that conscious processing can result in outcomes that conflict with our preferences, values, and desires. For example, Dijksterhuis et al. (2006) claim that conscious reasoning may help us meet our preferences when making simple decisions, but hinder our preferences when making more complex decisions. Dijksterhuis et al. focus specifically on consumer choices and conducted several studies where participants make either simple purchases (e.g. oven mitts, towels) or complex purchases (furniture, a home, art, a vacation). Dijksterhuis et al. found that participants who spent a significant amount of time thinking about the simple purchases were more satisfied with their purchases in the long term than those who made the purchase without much conscious reflection. In contrast, however, participants who spent a significant amount of time thinking about the *complex* purchases were *less* satisfied in the long term than those who made the complex purchase without much conscious reflection. Dijksterhuis et al. draw from their research the conclusion that "…it should benefit the individual to think consciously about simple matters and to delegate thinking about more complex matters to the unconscious" (1007).

Similarly, emerging empirical data suggest that automaticity can in some cases produce *morally* desirable outcomes more reliably than deliberation. Rand, Greene, and Nowak (2012) conducted a series of studies to explore the relationship between moral decisions and systems 1 and 2 discussed above. Specifically, Rand et al. wondered whether cooperative decisions are more likely to be produced by System 1 (automaticity), System 2 (deliberation), or both equally. In some of the studies conducted, participants were instructed to play a type of prisoner's dilemma, in which cooperation benefited everyone the most, but selfishness benefited the individual participant. To determine whether participants were being guided by System 1 or System 2, Rand et al. measured or controlled participants' response time. Shorter response time

was taken to indicate a quick, intuitive, automatic decision, while longer response time was taken to indicate a more deliberative, controlled, reflective decision. In each of the studies, participants who responded more quickly made more cooperative decisions (427). This suggests that system 1 more reliably produced cooperative decisions. In addition to the correlation between automaticity and cooperation shown in these studies, Rand et al. showed causation: subjects instructed to make quicker decisions made more cooperative decisions than those instructed to take their time making a decision (428). Additionally, subjects who were primed regarding the efficaciousness of intuition made more cooperative decisions than those who were primed regarding the efficaciousness of reasoning (428).[41] These studies suggest that in at least some cases, quick automatic processes, even when artificially induced, lead individuals to more morally desirable decisions—to be cooperative or giving, for example—than conscious reflection.[42]

Research on expert athletes and musicians also suggests that there are certain situations in which automaticity outperforms System 2 processing. Some performers, when asked how they play so well, respond: "I don't know; I just do." World-renowned jazz saxophonist Sonny Rollins articulates this kind of idea when describing his practice. When asked about his motivation for studying yoga and Eastern religious practices in relation to his musical practice, Rollins explains:

[41] See also Tomasello 2012.

[42] See also, for example, Small, Loewenstein, and Slovic 2007: "In a series of field experiments, we show that teaching or priming people to recognize the discrepancy in giving toward identifiable and statistical victims has perverse effects: individuals give less to identifiable victims but do not increase giving to statistical victims, resulting in an overall reduction in caring and giving. Thus, it appears that, when thinking deliberatively, people discount sympathy towards identifiable victims but fail to generate sympathy toward statistical victims" (143) and Zhong 2012: "The research presented here suggests that deliberative decision making may actually increase unethical behaviors and reduce altruistic motives when it overshadows implicit, intuitive influences on moral judgments and decisions. Three lab experiments explored the potential ethical dangers of deliberative decision making. Experiments1 and 2 showed that deliberative decision making, activated by a math problem-solving task or by simply framing the choice as a decision rather than an intuitive reaction, increased deception in a one-shot deception game. Experiment 3—which activated systematic thinking or intuitive feeling about the choice to donate to a charity—found that deliberative decision making could also decrease altruism" (1).

> When I play, what I try to do is to reach my subconscious level. I don't want to overtly think about anything, because you can't think and play at the same time — believe me, I've tried it (*laughs*). It goes by too fast. So when you're into yoga and when you're into improvisation, you want to reach that other level… *I'm not supposed to be playing, the music is supposed to be playing me*. I'm just supposed to be standing there with the horn, moving my fingers. The music is supposed to be coming through me; *that's when it's really happening* (NPR 2014, emphasis mine).

Here, Rollins suggests that he achieves excellence by shutting down his deliberative processes. While Rollins is able to articulate the importance of tapping into his subconscious, he would likely not be able to articulate any of the reasons for his various choices in an improv set. Most likely, if asked why he relied so heavily on the Dorian scale in a particular solo, Rollins would reply, "it just felt right." He might even suggest that he did not choose the Dorian scale, but that it chose him.

Another example illustrates that it is sometimes not only important to "turn off" one's conscious processes in the moment, but also to avoid trying to understand the intricacies and reasons for one's practice period. This kind of idea is becoming increasingly popular in sports psychology, specifically with regard to "Steve Blass Disease." Steve Blass was an excellent pitcher for the Pittsburg Pirates in the 1970s. However, for apparently random and inexplicable reasons, he started choking during games, leading to a downward spiral that resulted in a shameful early retirement and personal depression. One hypothesis is that Steve Blass could not stop thinking about his performance, which made him unable to perform. Ira Glass, from "This American Life," explains:

> In the decades since Steve Blass stopped pitching, researchers have been trying to figure out what to do about it and what is going on when a player falls apart like he did. And they believe that basically the problem comes down to thinking. When an elite athlete is at his or her best, when they're in the zone, their movements are automatic. They're not thinking about how their wrist turns, or their knee bends, or any of the other details. And when researchers bring athletes into the lab with a simulated batting cage, or a putting green, when they tell them to think about the mechanics of what they're doing, to notice where exactly the bat is moving when they're swinging, or how their elbow shifts when

> they're putting, the athletes—the overwhelming majority of them—start to choke. Thinking is the problem. (This American Life 2012, see also Weiss and Reber 2012)

Blass and other athletes are not only unable to articulate the norms or reasons behind their judgments, decisions, and actions, but suffer a cost when asked to do so.[43] Like Rollins, it seems that Blass achieved excellence by channeling his talent or the practice, not by calling upon conscious reflection. To engage in reflection during, before, or after the moment actually derailed his practice.

Notice that I am not simply suggesting that Rollins and Blass fail to engage in reflection at the moment of action. I am suggesting that their practices are not intellectually developed or informed in the way some might they should be. That is, it is not clear that their practices have involved deliberation about norms and reasons at any point. The development of their practices does not seem to involve internalization of norms through habit, but rather a turning off of system 2 processes so that natural talent or subconscious skills can be channeled. But does the inability to articulate his motivating reasons indicate an "unreasoning attachment" or inferiority to the saxophonist who can explain in detail the motivations and choices in her performance? I suggest not. Rollins and Blass's performances are not only accurate (similar to technically getting a moral judgment right), but also excellent, admirable, robust, deep, etc (pre-"Steve Blass Disease" at least). It would be a mistake to think that Rollins or Blass's "un-intellectually developed" skills are qualitatively lacking in some way. Instead, we should think of their practices as mature, robust, excellent, and worthy of praise similar to the saxophonist or pitcher who has a clear understanding of the rules and norms guiding her practice.

[43] Werkheiser (2014) even argues that asking for reasons reinforces social inequality by creating self-doubt in individuals from marginalized social groups.

Given the analysis in this section thus far, the following claims about automaticity and reasoning seem warranted.

- *There are two different processes/representations/systems that guide judgments, decisions, and behaviors*.
- *Both of these processes make important contributions in guiding our judgments, decisions, and behaviors.*
- *It appears that both processes are somewhat situationally constrained—meaning they function well in some contexts or situations, but not others.*
- *Both processes need and deserve further study.*

Now that the empirical question of whether automaticity and/or reasoning guide our moral judgments and decisions has been somewhat settled, it should become clear that the issues philosophers are really concerned about regarding Haidt's work are normative. If automatic processing alone does guide some/much/most of our moral judgment and decision-making, do our judgments and decisions lose normative credibility or force? I answer this question in the next chapter.

Conclusion

In this chapter, I have summarized how the philosophical conversation has developed since Haidt's 2001 essay. As I explained in Section I, there has been significant philosophical resistance to Haidt's thesis. In Section II, I attempted to unpack the nature and causes of this resistance (as I take them to be more complicated than mere intellectual disagreement). In Section III, I offered a report on our current general understanding of the nature and roles of automaticity and deliberation. My specific aim in this latter section was to show that the conversation up to this point has been somewhat oversimplified and stunted. The important question is not: which guides moral judgment and decision-making more, automaticity or deliberation? Instead, we should be focusing on questions like: when do automaticity and deliberation function well? How can we harness the power of automaticity and deliberation? We

already know they both guide judgments and decisions in different contexts and situations. And we know that sometimes automatic processing helps, while other times is hinders, just as deliberative processing sometimes helps while other times it hinders. We ought to be pursuing research, then, that aims to illuminate the boundaries, strengths and weaknesses, and potential for both automaticity and deliberation in moral judgment and decision-making. I pursue this project, focusing specifically on automatic processing, in Chapter 4.

Works Cited

Ambady, Nalini. 2010. "The Perils of Pondering: Intuition and Thin Slice Judgments". Psychological Inquiry. 21 (4): 271-278.

Bargh, John A., and Tanya L. Chartrand. 1999. "The Unbearable Automaticity of Being". American Psychologist. 54 (7).

Baumeister, Roy F., & Bargh, John A. 2014. Conscious and unconscious: Toward an integrative understanding of human life and action. In J. Sherman (Ed.), *Dual process theories of the social mind*. New York: Guilford

Bechara, Antoine, Damasio, Antonio R., Damasio, Hanna, & Anderson, Steven W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition,* 50 (1), 7-15.

Blair, Irene V. 2002. "The Malleability of Automatic Stereotypes and Prejudice". *Personality and Social Psychology Review*. 6 (3): 242-261.

Clarke, Steve. 2008. SIM and the city: Rationalism in psychology and philosophy and Haidt's account of moral judgment. *Philosophical Psychology* 21 (6): 799-820.

Craigie, Jillian. 2011. Thinking and feeling: Moral deliberation in a dual-process framework. *Philosophical Psychology* 24 (1): 53-71.

Dijksterhuis, Ap, Maarten W. Bos, Loran F. Nordgren, & Rick B. Van Baaren. 2006. On making the right choice: The deliberation-without-attention effect.(REPORTS). *Science, 311*(5763), 1005-1007.

Epstein, Seymour. 1994. "Integration of the cognitive and the psychodynamic unconscious". *American Psychologist*. 49 (8): 709-724.

Fine, Cordelia. 2006. Is the emotional dog wagging its rational tail, or chasing it?: Reason in moral judgment. *Philosophical Explorations* 9 (1): 83-98.

Gendler, Tamar. 2008. Alief in action (and reaction). *Mind and Language* 23:552-85.

Gollwitzer, Peter M. 1999. "Implementation Intentions". *American Psychologist*. 54 (7).

Haidt Jonathan. 2001. "The emotional dog and its rational tail: a social intuitionist approach to moral judgment". Psychological Review. 108 (4): 814-34.

————. 2003. "The Emotional Dog Does Learn New Tricks: A Reply to Pizarro and Bloom (2003)". Psychological Review. 110 (1).

————. 2004. "The Emotional Dog Gets Mistaken for a Possum". Review of General Psychology: Journal of Division 1, of the American Psychological Association. 8: 283-290.

————. 2007. "The new synthesis in moral psychology". Science (New York, N.Y.). 316 (5827): 998-1002.

————. 2010. "Moral Psychology Must Not Be Based on Faith and Hope: Commentary on Narvaez (2010)". Perspectives on Psychological Science. 5 (2): 182-184.

————. 2012. The righteous mind: why good people are divided by politics and religion. New York: Pantheon Books.

————. 2013. "Moral Psychology for the Twenty-First Century". Journal of Moral Education. 42 (3): 281-297

Horgan, Terry, and Mark Timmons. 2007. Morphological rationalism and the psychology of moral judgment. *Ethical Theory & Moral Practice* 10 (3): 279-95.

Kahneman Daniel. 2003. "A perspective on judgment and choice: mapping bounded rationality". *The American Psychologist*. 58 (9): 697-720.

Karremans, Johan C., & Aarts, Henk. 2007. The role of automaticity in determining the inclination to forgive close others. *Journal of Experimental Social Psychology, 43*(6), 902-917.
Kennett, Jeanette. 2012. Living with one's choices: Moral reasoning In Vitro and In Vivo. In *Emotions, imagination, and moral reasoning*, eds. Langdon, Robyn, and Catriona Mackenzie. New York, NY: Psychology Press.
Kennett, Jeanette, and Cordelia Fine. 2009. "Will the Real Moral Judgment Please Stand Up?" *Ethical Theory & Moral Practice*. 12 (1): 77-96.
Levy, Neil. 2006. "The wisdom of the pack". Philosophical Explorations. 9 (1): 99-103.
Lombrozo, Tania. 2009. "The Role of Moral Commitments in Moral Judgment". Cognitive Science. 33 (2): 273-286.
Moore, D., & Loewenstein, A. (2004). Self-Interest, Automaticity, and the Psychology of Conflict of Interest. *Social Justice Research, 17*(2), 189-202.
Musschenga, Albert W. 2008. Moral judgment and moral reasoning: A critique of Jonathan Haidt. In *The contingent nature of life bioethics and limits of human existence*, eds. Düwell, Marcus, Christoph Rehmann-Sutter, and Dietmar Mieth. Springer.
Narvaez, Darcia. (2008). The Social-Intuitionist Model: Some Counter-Intuitions. In W. A. Sinnott-Armstrong (Ed.), Moral Psychology, Vol. 2, The Cognitive science of morality: Intuition and diversity
NPR 2014. "Sonny Rollins: 'You Can't Think And Play At The Same Time'" May 3, 2014. http://www.npr.org/2014/05/03/309047616/sonny-rollins-you-cant-think-and-play-at-the-same-time
Paxton, Joseph M., and Joshua D. Greene. 2010. "Moral Reasoning: Hints and Allegations". *Topics in Cognitive Science*. 2 (3): 511-527.
Pizarro DA, and Bloom P. 2003. The intelligence of the moral intuitions: Comment on Haidt (2001). *Psychological Review* 110 (1): 193-6.
Railton Peter. 2014. "The affective dog and its rational tale: Intuition and attunement". Ethics. 124 (4): 813-859.
Rand David, Joshua Greene, and Martin Nowak. 2012. Spontaneous giving and calculated greed. *Nature*. 489 (7416): 427-30.
Rydell, Robert J., and Allen R. McConnell. 2006. "Understanding implicit and explicit attitude change: A systems of reasoning analysis". *Journal of Personality and Social Psychology*. 91 (6): 995-1008.
Saltzstein, H. D., and T. Kasachkoff. 2004. Haidt's moral intuitionist theory: A psychological and philosophical critique. *Review of General Psychology: Journal of Division 1, of the American Psychological Association.* 8: 273-82.
Sauer H. 2012. Educated intuitions. automaticity and rationality in moral judgment. *Philosophical Explorations* 15 (3): 255-75.
Sherman, Jeffrey W., Bertram Gawronski, and Yaacov Trope. 2014. *Dual-process theories of the social mind*. New York: The Guilford Press.
Sloman, Steven A. 1996. "The empirical case for two systems of reasoning". *Psychological Bulletin*. 119 (1).
Small, Deborah A., Loewenstein, George, and Slovic, Paul. 2007. "Sympathy and Callousness: The Impact of Deliberative Thought on Donations to Identifiable and Statistical Victims." *Organizational Behavior and Human Decision Processes* 102, no. 2 (2007): 143-53.

Smith, Eliot R., and Jamie DeCoster. 2000. "Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems". *Personality and Social Psychology Review*. 4 (2): 108-131.
Strack, Fritz, and Roland Deutsch. 2004. "Reflective and Impulsive Determinants of Social Behavior". *Personality and Social Psychology Review*. 8 (3): 220-247.
This American Life, 2012. "Own Worst Enemy." WBEZ Chicago Podcast. Transcript available at: http://www.thisamericanlife.org/radio-archives/episode/462/transcript
Tomasello, Michael. 2012. "Why Be Nice? Better Not Think about It." *Trends in Cognitive Sciences* 16 (12): 580-81
Turnball, Oliver H., Caroline H. Bowman, Shanti Shanker and Julie L. Davies. 2014. Emotion-based learning: insights from the Iowa Gambling Task. *Frontiers in Psycholo*gy 5:162.
Weiss S.M., and Reber A.S. 2012. Curing the dreaded "steve blass disease. *Journal of Sport Psychology in Action*, 3 (3): 171-181.
Wilson, Timothy D. 2002. *Strangers to ourselves: discovering the adaptive unconscious.* Cambridge, Mass: Belknap Press of Harvard University Press.
Zhong, Chen-Bo. 2011. "The Ethical Dangers of Deliberative Decision Making." *Administrative Science Quarterly* 56, no. 1, 1-25.

CHAPTER 3

INTRODUCTION

As discussed in the previous chapter, recent research in psychology and cognitive science suggests that some of our actions are guided by deliberation, some are guided by both deliberation and automaticity, and some are guided by automaticity alone. Is this latter fact a problem for our normative theories of moral judgment and decision-making? Do moral judgments, decisions, and actions that are guided solely by automatic processes have moral worth? Are we morally responsible for them? Does the work on automaticity challenge our agency? These are the questions I address in this chapter.

When an action has *moral worth*, it is not simply right or wrong, but warrants the actor praise or blame. Shawn is annoyed with the panhandler at his bus stop. In an attempt to get the panhandler to leave, Shawn reaches into his pocket and gives the panhandler what Shawn thinks is a fake bill that he and his kids had been playing with the day before. However, Shawn has mistakenly given the panhandler an actual $20 bill, with which the panhandler was able to buy several meals and a night at the shelter. This makes a significant impact in the panhandler's life and he is grateful for Shawn's charitable gesture and recognition. Here, much good is generated by Shawn's action,[44] though we would not think that he *ought to be praised* for his action. Remember, the action from his perspective was something like "give this guy fake money to get him out of my face." Given this, Shawn's action does not have moral worth (or, more specifically, does not have *positive* moral worth). We do not think that he deserves praise for his action. I follow Julia Markovits and Nomy Arpaly in this characterization of moral worth: "Morally worthy actions (the thought is) aren't just right actions—they are actions for which the

[44] Let's grant that the panhandler in this case having an extra $20 to meet his basic needs, which would have otherwise gone unmet, is a good thing, and hence giving him the $20 was a good moral action.

agent who performs them *merits praise*…. Morally worthy actions are ones that reflect well on the moral character of the person who performs them" (Markovits 2010, 203; emphasis original), and "The moral worth of an action is the extent to which the agent deserves moral praise or blame for performing the action, the extent to which the action speaks well of the agent" (Arpaly 2003, 69). Note that this is distinct from whether an action is right or wrong. Arpaly explains: "The extent to which an agent deserves praise or blame for her action depends in part on the action's *moral desirability*," where "moral desirability captures the right or wrongness and degree of the action" (69, emphasis original). Arpaly elaborates:

> Two actions that are equal in moral desirability may be of different moral worth. To give a simple example, two people may donate equal amounts of money to Oxfam, but one of them may do so out of concern for improving the state of the world, while the other does so purely at the urging of her accountant. Even if the two agents' charitable actions are equally morally desirable—both of them have done the right thing—it is not true that both agents deserve the same degree of praise. (69)

Arpaly further distinguishes between positive and negative moral worth, which correlate respectively to praise and blame (69). Here, I will focus primarily on cases of "positive" moral worth—that is, on circumstances where an agent may or may not deserve praise.[45] Hence, I will often use "moral worth" interchangeably with "warrants praise," though I take Arpaly's distinction to be useful and important. The central question for this chapter is whether actions that are not guided by deliberation have moral worth.

Another way of phrasing the question, one that engages a different subset of philosophical literature is: are agents *morally responsible* for their actions that are not guided by

[45] Note that I am not interested in the question of under what circumstances we *should* praise someone—for prudential reasons, we might praise someone whose action is right but does not possess moral worth because we would like to positively reinforce his action, for example. See Markovits and Arpaly for brief discussion of this distinction: "It is important to distinguish, in this context, between actions we have instrumental reasons to praise and actions that merit praise in their own right" (Markovits 2010, 203, footnote 4), and "The moral worth of an action is the extent to which the agent deserves praise or blame for the action, not the extent to which the agent should be morally praised or blamed for it" (Arpaly 2003, 71).

deliberation? When asking if agents are morally responsible for a particular action, we are again asking whether they deserve praise/blame for their action. Angela Smith states, for example:

> I interpret the fundamental question of responsibility as a question about the conditions for moral attributability, that is to say, the conditions under which something can be attributed to a person in a way that is required in order for it to be a basis for moral appraisal of that person. To say that a person is responsible for something, in this sense, is only to say that she is *open* to moral appraisal on account of it (where nothing is implied about what that appraisal, if any, should be). (Smith 2005, 238)

Smith's point here is that to say that someone is "responsible" for an action does not translate to them being blameworthy—e.g. "you are responsible for this mess because you left the puppy unattended to watch TV"—but rather means that they are open to some evaluation for their action.[46] That is, they are open to, or warrant, praise *or* blame. Hence, in terms of moral responsibility, the central question of this chapter is: are individuals morally responsible for their actions that are not guided by deliberation? Note that I have not said what gives an action moral worth or what makes one responsible for their actions. Explaining that is part of the project of this chapter. Thus far, I have simply tried to explain what is meant by "moral worth" and "moral responsibility": to claim that an action has moral worth is to say the agent is praise/blameworthy for it and to claim that one is responsible for his/her action is to say that he/she is open to praise

[46] Smith footnotes several other authors who endorse this kind of view of responsibility: "For similar interpretations of this central idea of moral responsibility, see Bernard Berofsky, *Freedom from Necessity: The Metaphysical Basis of Responsibility* (London: Routledge & Kegan Paul, 1987), p. 45: 'The core idea of moral responsibility for action A [is] the principle that A is proper to cite in a moral evaluation of the agent'; Justin Oakley, *Morality and the Emotions* (London: Routledge, 1992), p. 124: 'To be responsible for something is to be open to creditworthiness or blameworthiness for it, but whether one is actually creditworthy or blameworthy for it depends also on its goodness or badness (or rightness or wrongness)'; and T.M. Scanlon, *What We Owe to Each Other* (Cambridge, Mass.: Harvard University Press, 1998), p. 248: 'To say that a person is responsible, in this sense, for a given action is only to say that it is appropriate to take it as a basis of moral appraisal of that person. Nothing is implied about what this appraisal should be—that is to say, about whether the action is praiseworthy, blameworthy, or morally indifferent.'" Smith 2005, 238, footnote 2.

or blame for it. Given this, I will often mean the same thing when by "an action has moral worth" and an "agent is morally responsible for her action."[47]

For some readers, it may seem obvious that deliberation is not a necessary precursor for the moral worth of an action or an individuals' moral responsibility. One might think, "Deliberation rarely guides my moral judgments and decisions. Moral judgment and decision-making is typically very intuitive and reactive for me." While I hope some do find my thesis here intuitive, my target audience is readers who instead find more compelling the idea that deliberation plays an ineliminable role in good moral judgment and decision-making. I begin, then, by thinking in more detail about the hesitancy to assign moral worth or moral responsibility in cases where an individual's action does not involve deliberation of the System 2 kind that is intentional, conscious, propositional, logical, slow, and controlled.

SECTION I: THE NEED FOR DELIBERATION

The reasons why one might be hesitant to assign moral worth or moral responsibility in cases where actions are not guided by deliberation can be divided into two broad categories. First, people might resist attributing moral worth to purely automatic actions as we think about moral judgment and decision-making from the first-person perspective. From the first person-perspective, I might think that it is unfair to be held responsible for actions about which I have not reflected. For example, perhaps the very first time I meet my cousin's new boyfriend, I have a quick negative reaction of dislike. But after getting to know him more and taking some time to

[47] Julia Markovits notes that these may be in fact interchangeable: "In my discussion of other versions of the Right Reasons Thesis, I have take the thesis to be about *necessary* conditions for the moral worth of actions. Here I expand the thesis to state *necessary and sufficient* conditions for the moral worth of actions. There is, of course, considerable debate about what it takes for agents to be *morally responsible* for their actions, and it is very plausible that an action can have moral worth only if it is one for which the agent is morally responsible…" (Markovits 2010, 205, footnote 9; emphasis original)

deliberate about the ways in which he is great partner for my cousin, I come to have lots of affection for him. Were my cousin to find out about my first impression and take offense, I would argue that I couldn't control my first impression and that my real judgment is that developed through my deliberation, so it would not be fair for my cousin to hold me to my first impression, but rather only my developed judgment. Or one could imagine that I am traveling internationally and even though I have done significant homework about other social practices and norms, I end up in a very unusual situation and impulsively give a greeting that is offensive. I might think that it was unfair if I were blamed for this mistake, given that I had never imagined this possible situation and did not have time to think about what was the best way to respond, in the same way it would be unfair to hold me responsible for my patellar reflex.[48]

The same might be said for inverse cases where individuals are assigned *praise* that they do not think they deserve. On January 2, 2007 at a NYC subway platform, a young man began having seizures and fell onto the subway track seconds before the oncoming train arrived. Bystander Wesley Autrey, who had been standing on the platform with his two daughters, quickly reacted: he jumped onto the tracks and, realizing there was not enough time to raise himself and the man back onto the platform, pulled the young man into the center of the tracks, away from the third rail, and held him down while the train passed inches over their heads. Both men survived with minimal injuries. When asked about his heroism in interviews, Autrey simply said "I don't feel like I did something spectacular; I just saw someone who needed help...I did what I felt was right" (Buckley 2007). In one interview, he seems to need convincing that he did

[48] See, also, Smith (unpublished) describing the attraction of what she calls the "Conscious Self View": "The view derives much of its appeal from the fact that we often *feel* as if we are the most 'active' and most obviously the 'agent' of what we do when we consciously execute an intention that has been formed as a result of reflective deliberation" (14) and "This view seems to affirm and lend support to a compelling Kantian ideal of moral justice: the idea that, as Thomas Nagel put it, 'it makes no sense to condemn oneself or anyone else for a quality that is not within the control of the will'" (15).

something praiseworthy. When asked about what lesson he would like his daughters to take away, he redirects the question and instead gives thanks to the NYC Mayor and the Metropolitan Transit Authority for digging the ditches deep enough to fit two people. When the interviewer interrupts to press "But what about the lesson of jumping in and saving another person's life? That's amazing," Autrey replies "Well, I mean I did that like out of a split second reaction" (CBS News 2007). In these comments, Autrey implies that he responded *without thinking*,[49] and because of this he does not deserve praise or the title of "hero." He simply acted or perhaps *merely reacted*.[50] From his perspective, it appears to seem unfair, in the sense of "undeserved" or "unwarranted," to assign him praise.

An additional worry from the first-person perspective may be that moral judgments and decisions that we do not deliberate about—either in the moment or at some previous time—may not feel like they are genuinely ours. That is, we may not feel like we act as *agents* when we act without deliberation. Take the phenomenon of implicit bias. Joe takes himself to be a deeply committed egalitarian and actively promotes social justice in his community. He works for a non-profit organization that serves marginalized populations such as people of color, the poor and homeless, women and children, and the elderly. While Joe comes from a rather privileged background, he has studied the causes and impacts of marginalization and injustice in the US and has a deep understanding of the social, psychological, and moral issues that he encounters at work. He has chosen this particular job after much deliberation as a great way to promote equality.[51] Now, further imagine that Joe takes an Implicit Association Test (IAT) on his

[49] Autrey's experience and explanation here aligns with recent empirical research on the automatic nature of extreme altruism: "In two studies, we provided evidence that when extreme altruists explain why they decided to help, the cognitive processes they describe are overwhelmingly intuitive, automatic and fast (Rand and Epstein 2014, 4).
[50] It might be suggested here that Autrey is simply expressing modesty about his actions. However, in interviews his body language and tone indicate that Autrey genuinely thinks he did nothing remarkable.
[51] Let us presume his job does actually promote equality.

computer one day. The IAT prompts Joe to associate people of different races, genders, and ages with positive and negative words and objects (like good/bad, helpful/harmful, wallet/weapon). The test happens so quickly that Joe has no time to think about his judgments. The test results suggest that Joe, much to his surprise, harbors harmful stereotypical attitudes—for example, he more quickly and often associates Black faces with negative words and weapons, and more quickly and often categorizes women and the elderly as weak or incompetent. When Joe gets his results he is stunned. However, over the next few days, Joe notices what was outside of his awareness before: that he crosses the street when he sees a Black man at night, is more likely to question the testimony of the female patrons at his office, and is more likely to assume the elderly patrons are stubborn and noncompliant. These judgments and behaviors do not stem from Joe's System 2 reflective processes. In fact, when Joe engages in deliberation, he makes judgments and decisions that directly contrast those happening outside of his awareness. As Joe struggles to make sense of this dissonance he tells himself: "I am obviously worried about the impact these implicit attitudes may have on my judgments and decisions and want to change them. But, I don't even feel like they are *mine*. They are not my *true* judgments. My *real* judgment is that we are all equal and should be treated so. I express my agency when I choose to come to work each day to help others, not when these biases I didn't even know I had cause me to cross the street. I don't deliberate in that moment; that's not *me* making that decision." Here, Joe expresses the idea that automatic judgments and decisions that aren't guided by deliberation are not really ours, are not expressions of our agency and self-governance.[52] Hence, it would be odd to say that these kinds of automatic actions have moral worth or that we are morally

[52] For this kind of view, see Kennett and Fine 2009: "We think that a closer examination of the interaction between automatic and controlled reflective processes in moral judgment provides some counter to scepticism about our agency and makes room for the view shared by rationalists and sophisticated sentimentalists alike that *genuine moral judgments are those that are regulated or endorsed by reflection.* (78; emphasis mine)

responsible for them.[53] They seem instead to be the kinds of things that happen to us, not that we choose or govern.

In addition to worrying from the first-person perspective about the idea that one could be responsible for those judgments and actions that do not involve deliberation, one might also worry from the second and third person perspectives—meaning we may worry about how to evaluate others for their non-deliberative judgments and decisions. The first reason why we may think deliberation is needed for morally worthy actions is because it indicates that those actions are not *accidentally* right or wrong. For example, when you ask Amber why she is building a rocking chair, she says it is for her friend George who has a newborn on the way and Amber has been thinking for weeks about what would be a nice gift to give him and, remembering his love for woodwork and chairs, decided that this rocking chair would be exactly the kind of thing he would love and use often. In this scenario, there is no confusion about the rightness of Amber's action nor the moral worth. She is clearly praiseworthy for her action (presuming no unusual circumstances here like that she has been brainwashed or is plotting a defective chair to hurt George or something). Amber's deliberation over the matter makes clear that she is acting for admirable moral reasons—she cares for George and wants to give him a gift that he will use and love. In contrast, it is less obvious that one acts for moral reasons when simply responding or acting without thinking. Imagine that on the way to meet her new spouse Lena, Sam grabs a box of chocolates for Lena out of the impulse aisle. Sam doesn't really think about it and she does not have good reason to think that Lena will like the chocolates. Lena has never really expressed interest in chocolate and Lena even mentioned a few weeks ago that she was going to try to cut

[53] On a view like the one I have articulated here, we may be morally responsible for eradicating these implicit biases once we know we have them, but would not be responsible for having the biases or unknowingly acting on them (unless one deliberately choose at some previous time to have the biases, which is not the case for Joe and many others). There is growing consensus that we are morally responsible for addressing harmful implicit biases, but not for having them. See Kelly and Roedder 2008, Holroyd 2012, and Saul 2013, for example.

back on sweets. Given this, this gift will likely not be well received. Surprisingly, though, Lena recently decided to go back on sweets and so loves the gesture and the chocolates. Lena asks Sam why she purchased them and Sam says, "I don't know; I just felt like it. Spontaneity, I guess." I suspect many of us are less likely to praise Sam than Amber for the gesture. Some might think Sam does not deserve any praise at all. It appears as if Sam, unlike Amber, *accidentally* did a kind thing. Amber is praiseworthy, whereas Sam is lucky.

It should be fairly uncontroversial that individuals do not deserve praise or blame when their actions are *accidentally* right or wrong.[54] If the farm owner begins using more wind energy to increase his profits, he would not be praised if unbeknownst to him he reduced greenhouse emissions. If I asked your roommate when your birthday was and he texted March 21 instead of March 12, I would not be blamed for calling on the 21st instead of the 12th. In other words, accidentally right or wrong actions do not have moral worth or engage moral responsibility.[55] It might be that many people worry that actions that fail to be guided by deliberation are merely *accidentally* right or wrong given that we do not call to mind and evaluate various reasons and considerations. From the information we have, Lena could have just as easily been annoyed with Sam for the chocolates, reminding Sam that she was cutting back on sweets. It was really only by accident that Sam seemed to do the right thing, in contrast to Amber who intentionally does the right thing after deliberating and searching for the right reasons.[56]

---

[54] Again, this a point about what individuals *deserve*, not what would be prudential. There are certainly cases where an individual would not *deserve* praise or blame, but it would be prudential to give it nevertheless.

[55] For more on the idea that praise and blameworthy judgments and actions are not accidental, see Arpaly 2003 (referencing Kant's Prudent Grocer): "The grocer aims at increasing his profits. By lucky accident, it so happens that the action that would most increase his profits is also a morally right action. While this is all well and good, one is not inclined to give the grocer moral credit for this accident," (71) and Markovits 2010, 203, footnote 4.

[56] One might also think deliberation is important because it more often leads to the morally right judgment or decision. This is an empirical claim that is open to debate. While interesting and important, it does not add anything to the idea that an action cannot have moral worth unless grounded in the deliberation. This latter point is a normative claim that is a separate issue from the descriptive.

These examples and concerns indicate that we think it is important that people's (and our own) judgments and actions be non-accidentally connected to moral reasons which should be non-accidentally connected to the individual's motives and reasons. I do not want my neighbor to bring me soup (which appears to be a nice gesture) because she had leftovers and cannot stand waste; I want my neighbor to bring me soup because she cares about me and knows I have been busy lately. I do not want to be the kind of person who responds willy-nilly to my environment or whatever situation arises; I want to be directed, reasonable, thoughtful, and responsive in my judgments and decisions. It is often presumed, I suspect, that deliberation ensures that our motivating reasons correspond to the relevant moral reasons. Deliberating means that instead of impulsively paying for my friend's lunch simply because I felt like it (seeming to lack moral reasons entirely), I pay for her lunch because I am aware that she is in a difficult place financially and to pay for her lunch would not only be helpful but make her feel supported and cared about (in which case my decision is based entirely on the appropriate moral reasons). The latter is my acting toward the right person, to the right extent, at the right time, with the right motive, and in the right way. Without deliberation, our actions seem more accidental and capricious—in which case they do not have moral worth nor would it be appropriate to say we are morally responsible for them.

Note that deliberation on this view *need not occur in the moment*. Much of our moral judgment and decision-making will happen quickly and automatically. We would likely be unable to act at all if all of our judgments and decisions needed to be preceded in the immediate moment by deliberation. However, deliberation must have occurred at some previous point, if not in the moment, and our particular judgment or decision must be in some way guided or connected to that previous deliberation. Hence, the defender of the idea that deliberation must

contribution to our judgments and decisions simply means that judgments and actions that are not grounded in—cannot be traced back to—some kind of deliberation cannot be attributed to us in such a way that we would be morally responsible for them. That is, we can only evaluate automatic judgments and actions if they are built upon deliberation. Julia Driver explains the point well:

> ….on the most common interpretation of Aristotle's account of moral virtue, [for example] automatic behavior is compatible with virtue as long as the automatic behavior was the result of the right kind of "training up" of the agent. Indeed, this would be how one understands the method by which the regulation of the behavior occurs. The agent is, in the past, made aware of the relevant reasons and actually uses them to guide behavior. The appropriate training involves deliberation *at some point*. (2013, 288; emphasis original)

The point here is that it is no problem for the defender of deliberation if much of our judgment and decision-making automatically occurs in the moment as long as those automatic responses are developed and founded upon deliberation. Without such deliberation, we would not know for what reasons we are acting, which raises the concerns I have discussed in this section about fairness, agency, and mere luck. On this analysis, philosophers are right to be worried about the research suggesting that some of our judgment and decision-making is *purely automatic*, that it is not grounded in deliberation.[57]

III. Moral Worth Without Deliberation: Cases

In this section, I argue that the above idea that only actions guided by deliberation have moral worth or are morally evaluable is false. This is illustrated, in part, by the fact that there are cases in which judgments or actions cannot be traced back to deliberation, but we nevertheless

[57] One might argue that there is no need to be worried about "purely automatic" judgment and decision-making because it does not in fact exist. Even judgments and decisions that look purely automatic are grounded in deliberative processes. I think this view fails to appreciate that some of our judgments and decisions are entirely determined by automatic, System 2, processing, which I discuss in more detail in Chapter 2.

think the individual ought to be praised or blamed—meaning we think the action has moral worth. In this section, I will describe such cases and explain why deliberation is not a necessary condition for moral worth.

Above, I have presented cases that seem to show that deliberation *is* a necessary condition for moral worth. However, there are many contexts and situations in which we would hold someone responsible even if deliberation did not guide their judgment or action. The first kind of case is where one simply fails to deliberate. Imagine that you live in a country where an oppressive regime has been imprisoning and, you suspect, killing a particular population of citizens. You certainly feel terrible about the situation, but are also somewhat removed and are able to carry on with your routine without much disturbance. Once day, you hear a knock at your front door and answer it to find a family on your doorstep, clearly frightened. Before anyone can say a word, you usher the family in, having decided immediately that they will stay with you for their own safety. You do not deliberate about this decision. There is no discussion before you make your decision. You are presented with a moral choice and you make it immediately.[58]

Let us grant that your action is morally praiseworthy. Let us further grant that you did not deliberate in the moment. Perhaps, then, deliberation is *not* necessary for moral praise. Many, however, will suggest here that even if you did not deliberate in the moment about helping these

[58] Herlinde Pauer-Studer and David J. Velleman describe cases like these in "Distortions of Normativity" (2011). Pauer-Studer and Velleman explain that many individuals who helped Jewish refugees during the holocaust "failed or even refused to describe their actions in moral terms" in interview (2011, 353 footnote 41), giving responses like:
"You see a child, you see how, ... in the street, in the station, everything is refused, everything except death—and in the early morning light this child looks at you with his big eyes, with enormous eyes: what do you do? I did it, that's all" (Halter 1998, 74).
"I never spent my time asking why I did all that. I did it, that's all" (Halter 1998, 109)
"I decided nothing. A man knocked on my door. He said he was in danger. I asked him to stay, and he stayed till the end of the war" (Halter 1998, 108).
"I cannot give you any reasons. It was not a question of reasoning. Let's put it this way. There were people in need and we helped them . . . . People always ask how we started, but we didn't start. It started. And it started very gradually. We never gave it much thought" (Oliner and Oliner 1988, 216).
The interviewees appear either unable or unwilling to give reasons for their moral decisions. Many explicitly say they did not think about it. They simply responded. They just did. (Thank you to Colin Marshall and Sara Goering for pointing me to this kind of case).

refugees, you likely deliberated *in the past* and that deliberation guided your immediate decision when the family shows up on your doorstep. Hence, your action did involve deliberation and thus is morally praiseworthy.

However, this claim that you deliberated at a previous point needs further development. When did you deliberate? And what did that deliberation look like? Very likely, it wasn't "If someone shows up at my doorstep in need, I think I should, for moral reasons such as…, risk my life and bring them into my home" such that when the family showed up, you were immediately able to execute your previous deliberation. By the same token, it is likely not the case that "subway rescuer" Wesley Autrey had previously deliberated about what he would do if he was standing on the subway platform with his two daughters and someone fell into the tracks due to a seizure moments before the train arrived. This level of deliberation, at which one has imagined the specifics of the case and the various moral considerations that ought to influence one's decision, seems unlikely.

Perhaps, then, a more general deliberation has occurred at some previous point. Perhaps you (and Wesley Autrey) have deliberated about what it means to be a good person. You may have determined that being a good person requires being sensitive to others' suffering and a willingness to help fellow humans when they are in need, even if that may sometimes incur a cost or risk for you. It is this precise deliberation that allows you (and Autrey) to act quickly when unexpectedly faced with a life-threatening situation. Hence, it would be suggested, deliberation *did* play a role in your moral judgment and decision.

There are a few problems with this response, however. First, the pervasiveness of this kind of deliberation is unclear. Many people accept the norms and expectations of "good person" that are taught in their communities, religions, schools, or families. This means that it is entirely

possible that some people do good things without ever having deliberated about what it means to be a good person. But, even more problematic is that if this deliberation had in fact occurred, we would expect you and Autrey to be able to explain how that deliberation led to your action. However, data suggest that in many situations where individuals act generously, cooperatively, bravely, etc., the individuals are *not* able to explain what reasons or reasoning led to their judgment or action. Recall that when Autrey is asked about why he did what he did, he does not say: "Well, I've always believed that we should do whatever we can to help each other in desperate times," or "I was taught from a young age not to simply stand by when people are in need." Instead, he has a hard time explaining why he acted, ultimately concluding that he just did what he felt was right. This makes it hard to believe that it was Autrey's previous deliberation that led to his particular action in this moment. Even if Autrey's beliefs about what it meant to be a good person were internalized over time, we would expect him to be able to give some response or some analysis rather than his almost dumbfounded "uh, I just did what I felt was right."

It might be suggested, however, that I am expecting too much of a person in a situation like Autrey's. Perhaps it is understandable that Autrey does not articulate how deliberation affected his present choice. Maybe the norms of conversation discourage people from describing their deliberation, or the connections are not present in our mind in this clear and conscious way. To the latter point, if the connection is not accessible to the actor, it is not clear how the person was acting *for the reasons* that he/she had previously deliberated on. To the former point, it is quite natural for people to explain how previous deliberation led to their current choices. Take for example, Suzanne, who assisted Jewish refugees during the Holocaust. Suzanne attributes her actions during the Holocaust to her reflection on her values and morality:

> The effects of [Suzanne's parents'] teachings are reflected in Suzanne's assessment of herself as a young adult. She describes herself as being very honest, very independent, very capable of taking responsibility and risks, very helpful, and very ready to assert her convictions. While she did not see herself as religious, she did see herself as a "very moral person," which she attributed largely to her Protestant upbringing. While Suzanne attributes much of her morality to her parents, it was not developed out of a simple acquiescence to their expectations. She challenged them frequently, and they frequently had to discipline her. He mother depended primarily on reasoning to convince her daughter of the rightness and wrongness of certain behavior, and when this failed, "she just ignored me"…. Confrontation and disputation also marked her closest friendships. "We argued about everything," she says of her two closest friends, "life, death, religion, and so on." She insisted on screening all behaviors and comments through the prism of her own thinking and autonomously chosen principles. "I never just accepted what others said," she commented…Her sense of universal obligations to others…[stemmed from] her commitment to the principle that all persons are entitled to be free. (Oliner and Oliner 1988, 214-215)

In a case like Suzanne's, it is easy to see that previous deliberation led to her moral actions at a later point. However, Suzanne and Wesley Autrey's explanations and experiences differ significantly. Even some people who acted in the same ways and context as Suzanne describe their motivations and experiences very differently, saying things like: "I decided nothing. A man knocked on my door. He said he was in danger. I asked him to stay, and he stayed till the end of the war" (Halter 1998, 108) and "I cannot give you any reasons. It was not a question of reasoning. Let's put it this way. There were people in need and we helped them…People always ask how we started, but we didn't start. It started. And it started very gradually. We never gave it much thought" (Oliner and Oliner 1988, 216). The difference in these responses warrants an explanation when we are trying to parse out what motivates particular moral actions. Perhaps the best explanation is that Suzanne's action is in fact grounded in deliberation, whereas the other rescuers' and Wesley Autrey's are *not* grounded in deliberation.[59] If true, and if we still believe

[59] It might be suggested here that one cannot infer from the lack of articulation of moral reasons that the agents did not deliberate at some previous point. However, my point here is that some explanation must be given for the difference and a lack of deliberation explains this difference well. If individuals had deliberated at a previous point, we expect them not to be so dumbfounded when asked to explain their motivations.

that Autrey and the other rescuers deserve praise, it would follow that deliberation is not a necessary condition for moral worth.[60]

Another kind of case, discussed in depth by Nomy Arpaly (2003) involves deliberation that does not guide one's action. Arpaly presents a number of illuminating cases in her book; here, I will discuss Mark Twain's Huckleberry Finn.[61] Huckleberry Finn acts against his judgments produced by deliberation, but we nevertheless find him morally praiseworthy. As Huck's trip with the slave Jim nears the end, Huck begins to think heavily about the decision to turn Jim in to his "owner." Huck's explicit beliefs are that Jim is not a person, but rather is property, and the morally right thing to do is return this "lost/stolen/missing" property to its rightful owner. Here, Huck engages very clearly in "System 2" deliberation: he is able to access the beliefs and commitments he has been taught and follows them to their logical end: that he should turn Jim in. However, when the moment comes, Huck finds himself unable to turn Jim in. Huck is paddling away from Jim, who has just thanked Huck for his freedom, when Huck encounters two men on skiff looking for runaway slaves. When they ask Huck about Jim's skin color—Jim is some distance away—Huck genuinely thinks the right thing to do is turn Jim in, but he finds himself responding instead: "I didn't answer up prompt. I tried to, but the words wouldn't come. I tried for a second or two to brace up and out with it, but I warn't man enough—hadn't the spunk of a rabbit. I see I was weakening; so I just give up trying, and up and says: 'He's white.'" Here, Huck acts against the conclusion he arrived at through deliberation. His deliberation led him to conclude that he should turn Jim in, but Huck let Jim go free. Hence, it

[60] This might be a good spot to remind readers that my aim is not to show that deliberation *never* contributes to moral actions or moral worth. Instead, my claim is that it need not; that is, some actions that are not guided by deliberation have moral worth. I do believe deliberation lead Suzanne to her actions and I think she is morally praiseworthy for them. But I also believe that deliberation did not lead other rescuers or Wesley Autrey, and they are also morally praiseworthy.

[61] Some readers may be skeptical of the use of a fictional character here, but Arpaly is following a long tradition of using Huckleberry Finn as a case for theorizing about moral psychology. See, for example Bennett 1974, Driver 1996, Hill 1998, and Hursthouse 1999.

would be implausible to say that Huck's action was guided by his deliberation. Nevertheless, Huck does the right thing *and* we think he is praiseworthy for it.[62] This supports, then, the idea that deliberation is *not* a necessary condition for moral worth or moral responsibility. Arpaly expands upon the point:

> Huckleberry Finn is not an isolated case…. We all have friends, family members or acquaintances of this sort. We can all recall the likes of a student who, waving his copy of *Atlas Shrugged* in one's face, preaches that one should be selfish and then proceeds to lose sleep generously helping his peers. If philosophers were right in believing that…only actions derived by deliberation from one's moral principles are done for moral reasons, we would have to view these people as bad people who happen to have some fortunate inclinations in their makeup. More commonly, however, we treat these people as fundamentally good people who happen to be incompetent abstract thinkers. (2003, 78).

In short, if one aims to defend the claim that actions of moral worth or responsibility must be guided by deliberation, one will either have to show how deliberation guides Huck and others or show that the above individuals are not morally responsible for these particular actions. Either claim, I think, will be hard to defend.

In the previous section, I suggested that there are a variety of reasons for why we might think deliberation *is* a necessary condition—we think moral actions should be non-accidental, we might not think it fair to be held accountable for actions about which we do not deliberate, we might think that agency by definition requires deliberation. However, I have described handful of cases in which an individual's action is *not* guided by deliberation—not in the moment nor in the past—but we nevertheless are inclined to think of the individual as morally responsible.[63] This

[62] Bennett (1974) does not think Huck is morally praiseworthy, but I think his analysis is misguided and Arpaly presents a compelling argument to this effect (2003, 75-78).

[63] This argumentative strategy is common in moral psychology. See, for example, Nomy Arpaly: "In the next six sections, I will introduce a collection of cases in which phenomena that appear in real life with substantial frequency appear to be different, in striking ways, from the paradigmatic cases contemporary moral psychologists try to accommodate, and hence puzzling to contemporary moral psychology…. I hope to achieve at least one very significant goal, and that is to expand the domain of moral-psychological inquiry, enriching our philosophical discussion of human beings in all their complexity" (2003, 7-8), Angela Smith: "I will begin my defense of this alternative account of responsibility…by considering a few examples. These examples are meant to bring out the

suggests that, contrary to some of our intuitions about responsibility, deliberation is *not* a necessary condition for moral worth of a particular action. In the next session, I will further defend this idea.

III. Acting for Moral Reasons without Deliberation

In this final section, I will attempt to explain why, in cases like Autrey's, the Holocaust rescuers, and Huck Finn's, we assign praise even though it appears that deliberation did not guide their actions. Above in Section I, I identified several reasons for why we would be worried about the worth of purely judgments and actions, such as that we might be acting right or wrongly only accidentally or be unfairly held accountable. I suggested that the concerns stem from the fundamental worry that agents might receive credit (either in the form of praise or blame) for judgments and actions that are not connected to moral reasons—recall the farmer who uses wind power for profit, not to help with the environment. Deliberation is a clear way to connect our motivating reasons to moral reasons. For example: when faced with a difficult choice about end-of-life care about one's parents, it is presumed that deliberation helps ensure that the reasons guiding one's decision are morally *relevant*, such as wanting to minimize suffering, fairly distributing the obligations of care, respecting your parents' requests and wishes, etc. Deliberation helps ensure that we do not decide or act for morally *irrelevant reasons*, such as one's own fear of death, inclination to pinch pennies, affinity for the alliterative title of the

---

intuitive plausibility of [my alternative account], while at the same time casting doubt upon the claim that we ordinarily take choice or voluntary control to be a precondition of legitimate moral assessment" (2005, 240), George Sher: "The most straightforward way to show that all is not well with the searchlight view is to establish that it conflicts with many of our ground-level beliefs about who is responsible for what…. I will make this argument in much more detail in chapter 2. To do so, I will begin by presenting nine examples of agents who are unaware of the morally relevant features of their wrong acts or omissions, yet who still seem morally responsible for those acts or omissions" (2009, 17), and finally, Julia Markovits: "I hope to show that the Motive of Duty Thesis runs against the grain of some central and attractive elements of the Kantian approach to ethics and wrongly excludes some apparently admirable actions [discussed via examples] from having moral worth" (2010, 204).

caregiving organization, etc. These kinds of automatic, gut-reactions are morally irrelevant and it would be a failure of moral agency to base one's decisions upon them.

This suggests that being properly motivated by moral reasons rather than the deliberation in and of itself is the key for good moral judgment and decision-making. We think Amber's action of making the rocking chair for George has moral worth—is more than *merely* right—because she is so clearly motivated by the relevant moral reasons, not because she deliberated more than Sam. Similarly, Sam's gift to Lena lacks moral worth because it does not seem to be motivated by moral reasons, but rather an inexplicable impulse. If, in fact, what we care about is that particular judgments and actions are motivated by the relevant moral reasons, we see that the role of deliberation becomes secondary. Deliberation is necessary for an action to have moral worth only if deliberation is necessary for one to act upon moral reasons. In this section, I argue that deliberation is not necessary for acting upon moral reasons.

Nomy Arpaly argues for the connection between moral worth and the acting for the right reasons in chapter 3 of *Unprincipled Virtue: An Inquiry into Moral Agency*. She presents the following thesis (specifically about actions with positive moral worth):

> *Praiseworthiness as Responsiveness to Moral Reasons (PRMR)*: For an agent to be morally praiseworthy for doing the right thing is for her to have done the right thing for the relevant moral reasons—that is, the reasons for which she acts are identical to the reasons for which the action is right. (Arpaly 2003, 72)[64]

Recall the two individuals who donate to Oxfam, mentioned in the quote from Arpaly above. Let us grant that donating to Oxfam is morally right and that the reason for this is because it helps alleviate suffering in the world. The first person, Carlos, donates for precisely this reason. He

[64] In a later section, Arpaly modifies this thesis to accommodate for the fact that there may be different degrees of moral worth; for example, two actions may be praiseworthy, but one may be more praiseworthy than the other (84). Degree of praiseworthiness is not relevant for my purposes here, so I present this more simplified version of the thesis.

wants to alleviate suffering. The second person, Susan, donates because her accountant said to—presumably because it will be profitable for her in the long run. In this scenario, we would say that Carlos's reasons for donating—to alleviate suffering—are identical to the reasons for which the action is right. Susan's reasons for donating, on the other hand, are not only different from, but irrelevant to, the reasons for which the action is right. Hence, according to Arpaly's thesis, Carlos is praiseworthy, but Susan is not.[65]

Arpaly makes an important distinction here. She explains that the PRMR holds that an agent is morally praiseworthy when she has acted for the right reasons, *not* when she has acted for what *she believes* are the right reasons. She means to be contrasting her claim here to contemporary understandings of "acting from duty" or "for the sake of the fine" (73). She explains that one might think acting from duty or for the sake of the fine means acting out of a concern for what is morally right. However, Arpaly argues that acting out of concern for the right is neither sufficient nor necessary for an action to have moral worth. First, she argues that acting out of concern for the right cannot be enough to make an action praiseworthy because individuals can be confused about what is in fact right. She gives an example of an extremist who avoids killing Tamara because killing a fellow Jew is a grave sin (74). Here, Arpaly says that the extremist has both done the right thing (avoided killing Tamara) and done so *out of concern for the right* (because he believes that killing fellow Jews is wrong). However, that Tamara is *Jewish* is not what makes killing her wrong; instead, that she is an *innocent person* makes killing her wrong. Hence, the extremist does the right thing and does so out of concern for what is right, but he does not (as is required by PRMR) do the right thing *for the relevant moral*

[65] One might say that Susan *should* be praised, perhaps to reinforce her morally right behavior. Again, I am focusing on the normative, not prudential, reasons for praise. Additionally, I make no claim here about whether Susan should simply not be praised (normatively) or should be blamed. The point is simply that she is not praiseworthy as Carlos is.

reasons (illustrated by the fact that if Tamara had not been Jewish, the extremist would have killed her). Hence, concern for what is right is not *sufficient* for establishing the moral worth of an action. Furthermore, Arpaly argues that a concern for what is right is not *necessary* for an action to have moral worth. Here, she discusses in more detail Huckleberry Finn.

As I discussed above, Huck's deliberation does not guide his moral decision not to turn Jim in. It must be, then, that some other motive/reasons caused Huck's decision. Arpaly suggests that the reasons for Huck's decision not to turn Jim in were precisely those moral reasons that make the action right—that Jim is human. That Huck ultimately comes to see Jim as human, and not property, causes Huck to set Jim free even though Huck's explicit commitments tell him to do otherwise. Arpaly explains:

> Talking to Jim about his hopes and fears and interacting with him extensively, Huckleberry constantly perceives data (never deliberated upon) that amount to the message that Jim is a person, just like him. [Mark] Twain makes it very easy for Huckleberry to perceive the similarity between himself and Jim: the two are equally ignorant, share the same language and superstitions…. While Huckleberry never reflects on these facts, they do prompt him to act toward Jim, more and more, in the same way he would have acted toward any other friend…. Huckleberry is not capable of bringing to consciousness his nonconscious awareness and making an inference along the lines of "Jim acts in all ways like a human being, therefore there is no reason to treat him as inferior, and thus what all the adults in my life think about blacks is wrong."…. But…his reluctance [to turn Jim in] is to a large extent the result of the fact that he has come to see Jim as a person, even if his conscious mind has not yet come to reflective awareness of this perceptual shift. (77)

Again, Arpaly's point is that Huck finds that he cannot return Jim to slavery because he sees Jim as a person. This is the *relevant moral reason* for this particular *right action.* This is the morally significant reason even though, Arpaly says, Huck does not know or believe that it is the right reason. Huck does not act out of concern for the right. Huck thinks he acts because he is a "weak-willed" boy and cannot follow through on his commitments. We, however, know that Huck is wrong about this; we know that his action is guided by the relevant moral reasons.

Huck stands in contrast to the above extremist, who also does the right thing (not killing Tamara) but for the wrong reason (because Tamara is Jewish). Our inclination to evaluate the extremist's action as lacking moral worth and to evaluate Huck's action as praiseworthy, show that moral worth is actually anchored in an alignment with the right moral reasons, not one's belief or concern about what is right. These cases that Arpaly discusses show that the foundation of moral worth is not deliberation, knowledge about what is right or wrong, or a concern to do the right thing. What matters instead is that the individual's reasons for acting accord with the moral reasons making the action right or wrong. Arpaly further aims to show that we can act for the right reasons without knowing it, as in the case of Huck Finn.[66]

Julia Markovits (2010) expands upon Arpaly's PRMR thesis. In large part, Markovits's account closely resembles Arpaly's.[67] Markovits states her thesis initially as "According to what I will call the Coincident Reasons Thesis, *my action is morally worthy if and only if my motivating reasons for acting coincide with the reasons morally justifying the action*—that is, if and only if I perform the action I morally ought to perform, for the (normative) reasons why it morally ought to be performed" (Markovits 2010, 205; emphasis original). Markovits explains exactly what it means for one's motivating reasons to coincide with the reasons morally justifying the action—or, in Arpaly's terms, to act for the right reasons. According to Markovits, the "justifying reasons" of an action are those reasons that make the action right. For example,

[66] Nancy Snow (2010) makes a similar argument about virtue, arguing that one could act virtuously (genuinely, not simply coincidentally) without knowing it: "A virtue-relevant goal is a goal which, if the agent had it, would, under the appropriate conditions, result in the agent's performing virtue-expressive, that is, virtuous, actions…. If we accept the definition of virtue-relevant goal, it follows that an agent need not have the goal of being virtuous *tout court*, or even the goal of being virtuous in the sense of having a goal to have a specific virtue, such as patience or courage, to have goals which would result in performing virtuous actions. An agent might have the goal of being a good parent, good colleague, good nurse, good citizen, or good friend. Having these goals would result in the agent's performing the virtuous actions, since these roles carry associated virtues" (2010, 53). In other words, one might have goals outside of "acquire virtue X" or "be virtuous" that also result in the expression of virtuous actions.
[67] As I noted above in footnote 19, Arpaly adds a clause about *degrees* of moral worth to the version of the PRMR that I have presented here (Arpaly 2003, 84). Markovits takes issue with Arpaly's view on degrees of moral worth and I will again bracket the point.

your friend has asked you to come over and help with some babysitting while she is finishing work for an important deadline at her office. You may be motivated to say "yes" for several reasons. First, you might be motivated to go and help because you care about your friend, and her children, and you want to help in any way that you can. Alternatively, you might know that she made an amazing birthday cake earlier this week and that there will definitely be some leftover, so you want to stop by to have some cake. Going to watch the children is really an excuse to go have some of her delicious cake. In this latter scenario, your motivating reason (get cake) does not accord with the justifying reasons that make the action right (that babysitting helps out your friend). In the former case, however, your motivating reasons *do* coincide with the justifying reasons—you agree to go babysit because you care about your friend and want to help her out. Like Arpaly, Markovits holds that one need not be aware of her motivating reasons that coincide with the action's justifying reasons (222).[68]

Markovits elaborates on what a "motivating reason" is. In short, it is the reason for which one acts. By this, Markovits does not mean the causal reasons—facts about physics, interactions in the mind, being in a particular location at a particular time—but rather the reasons that motivate our choices and actions. Markovits writes:

> Other kinds of facts can also, of course, explain our actions, such as biochemical facts about our brains, or facts about how much sleep we've been getting, or how much coffee we've drunk; and these might also be called 'reasons'. But they could not be described as "*our reasons* for acting." I might snap at you because I have not gotten enough sleep lately, but this cannot be *my reason* for snapping at you—it can't be my motivating reason. My motivating reason will always be some fact on the basis of which I chose to snap. (221-222)

[68] As discussed above, one might not only lack awareness of her reasons, but could also be confused about her reasons. I might, for example, agree to babysit because I think I am properly motivated to be a good friend, but outside of my conscious awareness am actually motivated to go over to have some cake. Or vice versa: I might call to mind the cake and think of it as my motivating reason, though in fact I am actually motivated to help my friend because I care about her. In either case, what will matter for my account here (as well as Markovits's and Arpaly's) is the agent's *actual* not *perceived* motivating reasons.

Why did Huck choose not to turn Jim in? Because he had come to see Jim as a person and a friend. Why did the extremist not kill Tamara? Because he believed that he shouldn't kill Jewish people. Markovits suggests that we find one's motivating reason when we ask: "what were *her reasons* for acting as she did?" Not *the* reasons, but *her* reasons. Sometimes, we will be able to ask the individual this question to figure out her motivating reasons. However, we are often times deceived, confused, or dumbfounded about our reasons for acting (e.g. Huck Finn, Wesley Autrey) and hence an individual's report may not always help properly identify one's reasons (Markovits 221-222). Hence, Markovits claims that motivating reasons are not beliefs, but facts—facts about the reasons (which may, but not need, be beliefs) that led one to act as she did (221).

Markovits also explains what it means for a motivating reason to *accord* to a justifying reason. Must the motivating reason be: "I want to keep my promise because I cannot will a universal maxim in which people break promises?" or "I went into the burning house to gather the neighbor's child because I wanted to minimize harm and maximize pleasure?" No.[69] Yet, the motivating reason can also not be identified too generally or most people performing an action

[69] This helps explain why "ability to articulate one's reasons" should not be a requirement for moral responsibility/moral worth. People may not have access to the more fundamental reasons that make an action right. But they may nevertheless be properly motivated by the right reason. This makes the following points misguided: "Someone who is loyal will have learnt, from parents, teachers, and other role models, the value of loyalty and reasons for being loyal and for ceasing to be loyal. In a crisis, she may stick by a fellow worker suspected of misconduct. This is an immediate response—'He needed support'. But when asked *she will be able to give reasons*—reasons, for example, for sticking by fellow workers whose character you know, rather than believing the worst of them because of someone else's suspicions. *If she can think of no such reasons but just insists on solidarity without any reason, we move to thinking that this is not loyalty but unreasoning attachment"* (Annas 2011, 29, emphasis mine), "If someone tells you that a certain action would be wrong …you may ask why it would be wrong and if there is no satisfactory answer you may reject that advice as unfounded…moral judgements require backing by reasons. This is a point about the logic of moral judgement…" (Rachels 2003), and "A rational animal is aware of the grounds of her beliefs and actions, of the way in which perception tends to influence her beliefs or desire tends to influence her actions. She is able to ask herself whether the forces that incline her to believe or do certain things amount to good reasons to believe or do those things, and then determine what she believes and does accordingly" (Kosgaard 2010, 23).

would have the same motivating reason in cases where we would think we should assign different evaluations of moral worth. Markovits explains, for example:

> Both the self-interested reward-seeker and the altruist may rush into the burning house to save the trapped child. The fact that rushing in will allow them to rescue the child is, of course, a normative moral reason—indeed, a sufficient one—to do so. So both the reward-seeker's and the altruists' motivating reasons overlap in some sense with the normative reasons morally justifying their act. (227)

The motivating reason could be described as the same in these two cases: "why did she run into the house?" "Because she wanted to save the trapped child." Markovits explains, though, that this answer "to save the trapped child," does not capture the more *fundamental* motivating reasons for each actor:

> As anyone who has had to answer a persistent four-year-old's string of "why?"s can attest, the reasons for which we act, as well as the normative reasons justifying our actions, are often interrelated in complex ways. We generally act for chains of dependent motivating reasons, running from the less to the more fundamental….the reward-seeker's *fundamental motivating reason* [to collect a reward] is not also a significant normative moral reason to save the child—her chain of motivating reasons diverges in its more fundamental links from the chain of normative moral reasons, whereas the altruist's does not. (227, emphasis mine)

Here, Markovits explains that even if both the reward-seeker and the altruist rushed into the burning house to save the child, when we (like the four-year-old) ask "but why did they do that?" the fundamental reason for reward seeker will be "because she wanted money," whereas for the altruist is will be something like "because she wanted to help" or "because it was the right thing to do." "To save the child" is instrumental to the reward-seeker's noninstrumental goal to get the reward-money, whereas it is noninstrumental for the altruist.

Now that we have this idea of a chain of reasons, one might wonder how deep/far must we go down the chain to properly identify one's motivating reasons? Markovits claims that as soon as one hits a noninstrumental reason, that is far enough. Hence, in the case of the reward-seeker, the first reason offered might be "why did she go into the burning house?" "*To save the*

child." We could plausibly follow up: "Why did she want to save the child?" "*To get the reward money.*" We could keep going here: "Why did she want the reward money?" "*Because she is greedy.*" "Why is she greedy?" Etc. But we already have evidence that the reward-seeker's fundamental motivating reason (to get the reward money) does not coincide with the justifying reasons of rushing into the burning house. Hence, further pursuit is not necessary.[70] The same analysis applies to the altruist: "Why did she go into the burning house?" "*To save the child*." "Why did she want to save the child?" "*Because she wanted to help.*" Helping the child and the family *is* the normative reason for rushing into the burning house. Hence, we see here that the altruist's fundamental motivating reason coincides with the justifying reason. We could push further here, but it would not add to our existing analysis unless the more fundamental reason was in tension with the morally justifying reasons. "She wanted to help because she is compassionate*"* does not add anything to the evaluation of this particular action, though might tell us more about the altruist in other ways.

The altruist might even be a trained philosopher and when we ask why she wanted to help, the more fundamental reason might be "because she recognizes the centrality of care in our lives and is committed to caring for those who are vulnerable." It is, of course, fine if this reason is part of her motivational chain, but it is not necessary. More general moral principles like this could not be a required part of the chain of reasons. To make moral principles like this required for moral worth would exclude many cases that we think do in fact constitute moral worth. Markovits explains:

[70] Further pursuit might be interesting for other moral or psychological reasons. What to make of a person who would risk her life like that for reward, for example. We might want to know for other reasons about the causes of her greed. But this information, whatever it might be, will not change the moral worth of this particular individual action. If we find that the reward seeker is greedy because she grew up heartbreakingly poor, we would not suddenly praise her for saving the child from the burning house because she still did it for the wrong reason. We might feel sympathy and understanding, but would not assign praise.

> For the conditions established by the [Coincident Reasons Thesis] would be far too difficult to meet if they required that every agent recognize the kind of fundamental justification for her actions that even moral philosophers, who spend their lives arguing about such matters, struggle to identify. Let's assume, for the sake of argument, that Kant's formula of humanity provides the right fundamental account of what makes actions right or wrong: the most fundamental reason why performing some act is right is that doing so amounts to treating those people affected by our actions as ends in themselves, not mere means, and to respecting the unconditional value they have as rational beings. But surely an agent needn't be motivated by *this* fact in order for her act to have moral worth. Committed utilitarians, for example, are certainly capable of worthy action. (228)

Hence, Markovits is not suggesting that we must follow the chain of reasons all the way to the most fundamental reason. Instead, she claims that we are looking for motivating reasons that are *noninstrumental*. "To get the reward" is an instrumental reason for the noninstrumental reason "to promote my own interest", whereas "to help the child" for the altruist is a noninstrumental moral reason. In light of this, Markovits slightly modifies her Coincident Reasons Thesis: "We should understand the Coincident Reasons Thesis as pronouncing an action morally worthy whenever the *noninstrumental reasons* for which is it performed coincide with the noninstrumental reasons that morally justify its performance" (230; emphasis mine). In other words, my action is praiseworthy when my motivating reasons—which will be noninstrumental, but not necessarily most foundational—coincide with the moral reasons that in fact make the action right. (Conversely, my action will be blameworthy when my motivating reasons coincide with the moral reason that make the action wrong, or when my reasons fail to coincide with the right reasons).

One might press here: why exactly does the "coincidence" of reasons constitute moral worth? I raised the worry above that sometimes a person might do the right thing accidentally: the business owner only incidentally benefits the environment because she wants to increase her profits. I described the problem with such a case as the person failing to do the right thing for the

right reasons—being motivated by greed instead of environmental good. I then argued that an action is morally praiseworthy when a person does the right thing for the right reason, even if she doesn't realize she is acting for the right reason—Huck is praiseworthy even though he doesn't know he acted for good moral reasons. However, it may seem in this latter case that because the right thing is done without conscious deliberation, it is also in a sense "accidental." For something to be *non*-accidental, it is commonly thought that it must be deliberate, expected, planned, intentional, designed, decided, premeditated, etc. For example, to say that injuries are non-accidental is to say that they were caused deliberately or decided. Thus, one might argue, even though Huck does the right thing *for the right reasons*—that is, his motivating reasons coincide with the justifying reasons of the action—his reasons are nevertheless only accidentally connected to the right action because they are not deliberate or intended to be right. Huck does not act on these reasons because he thinks they are right; his lack of deliberation about the right reasons prevents him from recognizing them as such. Given that I have granted here that accidentally morally right or wrong actions do not have moral worth, if cases of "coinciding moral reasons" are also accidental, they would then lack moral worth.

The reason why motivating reasons are not accidental, and thus are taken to indicate moral responsibility for the relevant action is because motivating reasons stem from our attitudes, commitments, and values. For example, the motivating reason for sending her brother a get well card was that Maya wanted to make him momentarily happy in a time when he is not feeling well. This motivating reason coincides with the justifying reasons that make the action morally right. The motivating reason further reflects Maya's attitudes, commitments, and values.[71] Her motivating reason indicates that she loves and values her brother as a person, that

[71] Here, I am invoking Angela Smith's "Rational Relations View," which holds that "When we praise or criticize someone for an attitude [or action]…it seems we are responding to certain judgments of the person which we take to

she values their relationship, that she values his happiness, that she is committed to being a good sister and perhaps a thoughtful person more generally, that she (at that moment) inhabits an attitude of care and selflessness, etc. That Maya's motivating reason coincides with the justifying reasons for sending the card is not merely accidental or morally irrelevant (in the way the prudent grocer's motivating reason to increase profits is).

As Maya need not be consciously aware of her motivating reason, she also need not be aware of her attitudes, commitments, or values that are reflected in this motivating reason. Lack of awareness of our attitudes, commitments, or values does not make them any less ours or make us any less answerable for them. Angela Smith explains, in arguing that we are morally responsible for passive states like omissions and involuntary reactions:

> …the [evaluative] judgments I am concerned with are not necessarily consciously held propositional beliefs [e.g. "I am committed to being a thoughtful person."], but rather tendencies to regard certain things as having evaluative significance. These judgments, taken together, make up the basic evaluative framework through which we view the world. They comprise the things we care about or regard as important or significant. "Judgments" in this sense do not always arise from conscious choices or decisions, and they need not be consciously recognized by the person who holds them. Indeed, these judgments are often things we discover about ourselves through our response to questions or to situations. For example, I may not realize, until I am faced with a choice, that I value the intellectual freedom and autonomy associated with a career in academia more highly than the economic rewards and benefits associated with a career in law. Or I may discover in some situation that I care more about being liked by others than I do about standing up for my moral principles. Although I may never have consciously entertained these evaluative judgments, I see that they are correctly attributable to me in virtue of my own responses to the situations I confront. (Smith 2005, 251-252).[72]

Perhaps Maya has recently reconnected with her brother after they were separated as young children and is thus surprised to find that she cares for her brother as deeply as she does. Even so, we would still say she is kind and morally admirable for sending him a get-well card. This is

---

be implicit in that attitude [or action], judgments for which we consider her to be directly morally answerable" (Smith 2005, 251); "…what makes an attitude [or action] "ours" in the sense relevant to questions of responsibility and moral assessment is that…it reflects our own evaluative judgments or appraisals" (237). See also Smith 2007, Smith 2008, and Smith 2012.

[72] For more on the "dawning" of our implicitly held attitudes, beliefs, and values, see also Arpaly 2003, 54-56.

also perhaps a way of explaining Huckleberry Finn's situation: even though he thinks he takes Jim to be nothing more than property, when pushed to make a decision about turning Jim in, he "finds" that he sees and values Jim as something more.

One might further press that there is still a sense in which coinciding motivating and justifying reasons could be accidental: namely in cases where one does not choose her evaluative judgments (attitudes, commitments, and values) that underlie the motivating reason. Imagine, for example, that Alejandra, who is typically misanthropic, is brainwashed to be thoughtful and giving or that Terrance has been brain damaged in such a way that he is unable to properly consider his own needs and is therefore necessarily selfless in his actions. If a case arose where the motivating reason for Alejandra and Terrance was to help Anthony, a motivating reason that would coincide with the justifying reason for doing the action, we would nevertheless be hesitant to assign moral responsibility to Alejandra and Terrance for their action—that is, we likely wouldn't view them as deserving of moral credit. Doesn't this speak against the thesis that "*my action is morally worthy if and only if my motivating reasons for acting coincide with the reasons morally justifying the action*—that is, if and only if I perform the action I morally ought to perform, for the (normative) reasons why it morally ought to be performed"?

Markovits does not consider unusual cases like these, but can respond to them by explaining that "to help Anthony" is not genuinely Alejandra and Terrance's motivating reason. It is the reason that motivates Alejandra and Terrance, but there is a way in which it cannot be said to be their motivation. In other words, "my" matters in "my motivating reasons."[73] Smith elaborates on such a point:

[73] This might not be a satisfying response from Markovits nor is it necessarily the response she would give. Another possible response is to simply qualify the Coincident Reasons Thesis with something like: "barring any sort of alien or artificial alteration of my motivations…."

> …it seems that an attitude "implanted" by a mad scientist, or one induced through posthypnotic suggestion, would also fail to meet the rational relations condition I have described…Since these attitudes are, by hypothesis, detached from a person's own rational assessment, it would be inappropriate to demand that she defend them, or to take them as a basis of rational or moral criticism. They do not really "belong" to her in a way that would make it possible to draw an inference about the evaluative judgments she accepts. (2005, 261-262).

Smith's point here helps to see that it is somewhat inaccurate in the above cases, where one is brainwashed or brain damaged, to say that the motivating reason under evaluation is *Alejandra's* or *Terrance's*, even though the reason does seem to motivate their action.

A more tricky case would be one where I have come to acquire problematic motivations, attitudes, commitments or values unintentionally, but not necessarily by implantation or being brainwashed. Smith explores these less obvious cases:

> Consider a person, Abigail, who is raised in a family or community that is deeply racist, say, or religiously intolerant. It would not be at all surprising for Abigail to develop evaluative tendencies and corresponding attitudes in line with those she sees operative in her family or surrounding community. As an adult, her attitudes may continue to reflect the vicious evaluative judgments thus formed in her childhood. Now contrast this person to another, Bert, who was raised in a loving and tolerant home and community, but who later in life reflectively comes to adopt racist and intolerant values. Bert's attitudes, like Abigail's, now reflect these vicious evaluative judgments, though unlike Abigail he formed these evaluative dispositions after he reached the age or rational maturity. What should we say about these two people? (Smith 2005, 267)

Here, Smith gets at the point that it seems to matter how one acquires one's values, commitments, and beliefs. She then elaborates.

> First of all, I think we can acknowledge that Bert is, while Abigail is not (or is not solely), responsible for becoming a racist or intolerant person. The seeds of Abigail's racism-intolerance were planted well before she was capable of rationally reflecting upon these evaluative commitments, while Bert's racism-intolerance is the result of his own mature reflective endorsement. So if we are asking whether each of these people is responsible for becoming racist or intolerant, our verdict might be mixed. But this question of responsibility (namely, the responsibility one has for becoming a certain kind of person) must be distinguished from the question of one's responsibility for the attitudes one in fact holds. In order to regard an attitude as attributable to a person, and as a legitimate basis for moral appraisal, we need not also claim that a person is responsible for becoming the sort of person who holds such an attitude. That is a separate question

> according to the view I am putting forward. What matters, according to the rational relations view, is that the attitude is in principle dependent upon and sensitive to the person's evaluative judgments. If a person continues to hold the objectionable attitude even after she has reached rational maturity, it is reasonable to attribute that attitude to her and to ask her to defend the judgment it reflects. It is worth noting here that if a person responded to such a demand by saying, "I am not responsible for my attitude—I was just raised this way," we would not feel compelled to withdraw our criticism. Citing the origin of one's attitude is irrelevant when what is in question is its justification. (Smith 2005, 267-268)

According to Smith, one would be morally responsible even in cases where she did not choose or deliberate about her fundamental attitudes, commitments, or values. Hence, the Coincident Reasons Thesis—my action is morally worthy if and only if my motivating reasons for acting coincide with the reasons morally justifying the action—applies even in cases where one has not chosen the values, commitments, and beliefs that cause one's motivating reasons.

In sum, in this section, I have provided defense for the idea that individuals can be morally responsible for judgments and decisions made outside of deliberation. This includes not only deliberation in the moment (which many take as obviously true) but deliberation at any point. Theorists like Arpaly, Markovits, and Smith help explain why deliberation is not a necessary condition for moral worth/moral responsibility: what matters is not that one deliberates, but that their motivating reasons for acting accord with the moral reasons that make the action right. Arpaly helps clarify that this does not simply mean that one is morally motivated—recall the extremist who did not kill Tamara because of his moral commitment and motivation. What matters is that one has the *right* moral motive. Markovits parses out what it means to *have the right motive*. The right moral motive is one that accords with the appropriate moral reasons at the noninstrumental level. Hence, one's motive need not be "I did X because I believe I should maximize happiness and minimize suffering." "Because I wanted to help" is sufficient. Deliberation may in some cases help one arrive at their motivating reasons—perhaps

when I am in the position of trying to decide whether to keep my plans for the weekend—but it need not. Sometimes, people arrive at the appropriate motivating reasons without any deliberation. Smith's work helps explain why are still responsible for these motivating reasons that we do not choose or occur outside of our awareness—because they reflect our values, commitments, and judgments. In the next chapter, I will discuss in more detail the significance of such values, commitments, and judgments.

Conclusion

I started this chapter off by introducing the cognitive distinction between System 1 and System 2. Much research in psychology and cognitive science suggests that many (at least more than previously assumed) of our judgments and decisions are in large part, sometimes exclusively, driven by System 1. In Section I, I tried to unpack the concerns that may guide the claim that System 2 must play a role in a judgment or action for it to have moral worth. I suggested that the foundational concern was that judgments and actions guided solely by System 1 lack the appropriate reasons or motivation. In Section II I argued that there are many cases where someone's action or judgment does not seem to be guided by deliberation, but we nevertheless assign moral worth to the action or moral responsibility to the agent. And in Section III, I attempted to show how one could be motivated by the relevant moral reasons, given one's action moral worth or making one morally responsible, without having deliberated. This suggests that System 1 can play an important normative role in our moral lives and thus it is worth further investigating how System 1 works, what the boundaries of System 1 are, and how exactly it

influences—both positively and negatively—our moral judgments and decisions. I provide an account that answers these questions in the next chapter.[74]

[74] Smith also indicates that the next move in this literature is to investigate the relationship between the theoretical and empirical: "The debate [between Shoemaker and Smith] might be understood as a debate over whether our more spontaneous 'moral intuitions' [as described by Haidt] can really be said to embody evaluative judgments. I don't think we (yet?) have convincing experimental evidence that can settle this question one way or another" (Smith 2012, 582, footnote 11). My distinguishing feature might be focusing on positive/praiseworthy cases.

Works Cited

Annas, Julia. 2011. *Intelligent virtue*. Oxford [England]: Oxford University Press.

Arpaly, Nomy. 2003. Unprincipled virtue: an inquiry into moral agency. Oxford: Oxford University Press.

Bennett, Jonathan. 1974. The Conscience of Huckleberry Finn. *Philosophy*, 49 (188): 123-134.

Block, Gay, Malka Drucker, Cynthia Ozick, and Harold M. Schulweis. 1992. *Rescuers: portraits of moral courage in the Holocaust*. New York: Holmes & Meier.

Buckley, Cara. 2007. Man Is Rescued by Stranger on Subway Tracks. *New York Times*. January 3, 2007. http://www.nytimes.com/2007/01/03/nyregion/03life.html?_r=0

CBS News. 2007. Subway Savior Speaks. January 4, 2007. http://www.cbsnews.com/videos/subway-savior-speaks/

Driver, Julia. 2013. "Moral expertise: judgment, practice, and analysis". Social Philosophy and Policy. 30 (1-2): 280-296.

Driver, Julia. 1996. "The Virtues and Human Nature." In Roger Crisp (ed.) *How Should One Live? Essays on the Virtues.* Oxford: Claredon Press. 111-130.

Halter, Marek. 1998. *Stories of deliverance: Speaking with men and women who rescued Jews from the Holocaust*. Chicago: Open Court.

Hill, Thomas. 1998. Four Conceptions of Conscience. *NOMOS* 60, 13-52.

Holroyd, Jules. 2012. "Responsibility for Implicit Bias." *Journal of Social Philosophy* 43 (3): 274-306.

Hursthouse, Rosalind. 1999. *On virtue ethics*. Oxford: Oxford University Press.

Kelly, Daniel, and Erica Roedder. 2008. "Racial Cognition and the Ethics of Implicit Bias." *Philosophy Compass* 3 (3): 522-40.

Korsgaard, Christine. 2010. "Yes, if…" in *Does Moral Action Depend on Reasoning?* John Templeton Foundation. Available at: http://www.templeton.org/reason/Essays/korsgaard.pdf

Markovits, Julia. 2010. "Acting for the Right Reasons". Philosophical Review. 119 (2).

Oliner, Samuel P., and Pearl M. Oliner. 1988. *The altruistic personality: rescuers of Jews in Nazi Europe*. New York: Free Press.

Rachels, James. 1993. *Subjectivism*. In: Singer P (ed) A Companion to ethics. Blackwell, Oxford. 432–441

Rand, David G, and Ziv G Epstein. "Risking Your Life without a Second Thought: Intuitive Decision-making and Extreme Altruism." *PloS One* 9, no. 10 (2014): E109687.

Saul, Jennifer. 2013. "Implicit Bias, Stereotype Threat, and Women in Philosophy." In *Women in Philosophy*, Oxford University Press.

Sher, George. 2009. *Who knew?: responsibility without awareness*. New York: Oxford University Press.

Smith, Angela M. 2005. "Responsibility for Attitudes: Activity and Passivity in Mental Life". *Ethics*. 115 (2).

Smith, Angela. 2007. "On Being Responsible and Holding Responsible". *The Journal of Ethics*. 11 (4): 465-484.

Smith, Angela M. 2008. "Control, responsibility, and moral assessment". *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*. 138 (3): 367-392.

Smith, Angela M. 2012. "Attributability, Answerability, and Accountability: In Defense of a Unified Account". *Ethics*. 112 (3).

CHAPTER 4

Introduction

In the previous chapters, I have attempted to describe key points in the conversation on moral judgment and decision making in the last 15 years. The starting point for my analysis is Haidt's 2001 critique of rationalism in philosophy and psychology. Haidt argues there, and elsewhere, that much of our moral judgment and decision-making is caused by affectively laden gut-reactions/intuitions and moral reasoning is often a tool used for post hoc rationalization of our judgments and decisions after they have already occurred. In Chapter 1, I explained his critiques and own model (Social Intuitionism) in detail. In Chapter 2, I reviewed various objections to Haidt's work. The primary thrust of these objections was that Haidt's thesis does not speak to our experiences—Haidt claims we rarely engage in moral reasoning, however it appears (from our own observations and some empirical research) that we engage in genuine deliberation about moral issues often (or at least not rarely). It is concluded, then, that Haidt has seriously underestimated or downplayed the role of genuine deliberation in our moral lives. However, as I argue in Chapter 2, the dialogue as currently framed reaches a stalemate—there is no way to adjudicate the conflicting claims of Haidt and his respondents. I explain why this stalemate arises and then reframe the conversation so that more productive conversation can occur about the nature and norms of automatic moral judgment and decision-making—specifically emphasizing the intelligence and potential of automatic processes. However, even if empirical research suggests that automatic processes are much more dynamic and responsive than previously described, it may still be argued that deliberation *should* play a central role in our moral judgment and decision-making; judgments and decisions not based on deliberation lack moral worth (in other words, we are not morally responsible for them). I argued in Chapter 3 that this

idea is not compelling—deliberation is not a necessary[75] condition of moral worth/moral responsibility. In this Chapter, I build upon my analysis by explaining in more detail exactly how automaticity serves us well as moral judgment and decision-makers. I present and defend an account of what I call "value-guided automaticity."

Value-guided automaticity is automatic processing—quick and outside of conscious awareness—for which values are the primary input that guide moral judgment and decision-making. This process happens without deliberation: our automatic system draws from our values without any deliberate thought or conscious awareness. Furthermore, I will focus on instances where the values themselves are not acquired via deliberation, but instead via automatic processes. Thus, deliberation may not play a role in generating the particular judgments or decisions that are produced by value-guided automaticity, yet nevertheless, such judgments and decisions are normatively evaluable, meaning it is prima facie appropriate to evaluate the actor as either praise- or blameworthy. Moral judgments and decisions lacking guidance by conscious reflection may nevertheless be grounded in an individual's implicitly held values. Implicitly held values would be values—such as friendship, loyalty, and generosity—that an individual has come to hold and continues to have without his or her awareness. For example, in the story of Huckleberry Finn, it could be said that Huck implicitly holds the value of friendship. None of his moral teachers have explained the meaning or importance of friendship, so it would not be said that he has acquired this value through deliberation, but it is nevertheless clear from his

[75] I also suspect that deliberation is not a sufficient condition for moral responsibility, though this is a question I bracket for future research. For now, let me mention that some beings such as psychopaths and robots may have excellent deliberative skills and engage in deliberation about moral questions, yet nevertheless lack moral agency or responsibility. In chapter 5, I discuss the idea that one must have the *capacity* to distinguish between right and wrong to be appropriately considered morally responsible. In one sense, the psychopath and robot may have the capacity to distinguish between right and wrong—they may be able to know and report the categorical imperative or harms principle, for example. But there is another sense in which I think the psychopath and robot do not *really know* or *really understand* the difference between right and wrong. Again, this is an issue that warrants more attention than I can give here; I simply want to raise the possibility that deliberation is neither necessary nor sufficient for moral responsibility.

behaviors and reactions that friendship is very important to him, though perhaps not explicitly. He might not, for example, recognize that friendship is important to him or be able to articulate the role that friendship plays in his moral decisions.

While Huck's reflective processes tell him to turn runaway slave Jim in to his "master," something "tugs" at him in such a way that he is not able to do it. Like many of the people described above, Huck cannot consciously access or articulate the reasons that ultimately lead him to let Jim go free (in fact, Huck gives false reasons for his decisions, namely that he is a wicked and weak-willed boy). There has been much debate about the (ir)rationality of Huck's decision and the influence of emotion on moral judgment and decision-making. According to those who think deliberation is central to good moral judgment and decision-making (in the words of Nomy Arpaly): "[Huck] is no more morally praiseworthy for helping Jim than a seeing-eye dog is praiseworthy for its helpful deeds" (Arpaly 2003, 9). However, if we ground Huck's moral decision in an implicitly held value, such as friendship, his action is in fact more substantive than a dog's.[76] The capacity to hold values and act upon them (intentionally or not) separates Huck from mollusks, babies, and thermostats. And it does seem to be the case that the value of friendship drives Huck's inability to turn Jim in. As Huck is paddling to shore, away from Jim, to find someone to report Jim to, the following encounter happens:

> As I shoved off, [Jim] says: "Pooty soon I'll be a-shout'n' for joy, en I'll say, it's all on accounts o' Huck; I's a free man, en I couldn't ever ben free ef it hadn' ben for Huck; Huck done it. Jim won't ever forgit you, Huck; you's de bes' fren' Jim's ever had; en you's de *only* fren' ole Jim's got now." I was paddling off, all in a sweat to tell on him; but when he says this, it seemed to kind of take the tuck all out of me. I went along slow then, and I warn't right down certain whether I was glad I started or whether I warn't. (Twain, 235)

[76] Note that I am very open to the possibility that some animals are morally responsible given the robustness of their capacities to care and hold values. I discuss this idea more, though still briefly, in the latter section of this chapter.

Moments later, two men on a skiff approach Huck looking for runaway slaves. When they ask what color the man is on the raft down the way (the raft Jim is on), Huck pauses: "I didn't answer up prompt. I tried to, but the words wouldn't come. I tried for a second or two to brace up and out with it, but I warn't man enough—hadn't the spunk of a rabbit. I see I was weakening; so I just give up trying, and up and says: 'He's white.'" Here, Huck does the right moral thing. He does not turn Jim in, giving Jim freedom. And it appears that Jim's unintentional appeal to Huck's value of friendship is the primary trigger for Huck's decision not to turn Jim in. However, Huck fails to see the role that this value plays in his decision—he fails to see that he has the value altogether. The fact that Huck's decision is not dependent upon deliberation about moral reasons or norms does not mean it is wrong, superficial, or lacking in agency. Instead, as Arpaly explains:

> Twain takes Huckleberry to be an ignorant boy whose decency and virtue exceed those of many older and more educated men, and his failure to turn Jim in is portrayed not as a mere lucky accident of temperament, a case of fortunate squeamishness, but as something quite different….Twain obviously sees him as praiseworthy in a way that he wouldn't be if he were merely acting out of some atavistic mechanism…[Huck] is not treated by his creator as if he were acting for a nonmoral motive, but rather as if he were acting for a moral motive—*without knowing* that it is a moral motive. (Arpaly 2003, 9-10; emphasis original)

My suggestion is that Huck's response is not simply the result of "unreasoned attachment," but an example of how automatic mental processes can lead to judgments and decisions by drawing upon our implicitly held moral values without our awareness.[77]

In Section I, I will define "values" and give an overview of empirical work on the expression of our values—describing "value-guided automaticity." I will also describe in more detail how we acquire our values, drawing from psychological research on topics such as

[77] In fact, Huck seems to be a case of Frankfurt's "volitional necessity" insofar as it seems like he cannot do otherwise (Frankfurt 1988, 86; Shoemaker 2003, 104; Sripada 2015, footnote 35 cf Wolf 1993). I discuss this point in more detail below.

personality development, the acquisition of implicit biases, and developing concern for others. In Section II, I will explain how such value-guided automaticity meets the normative requirements of moral worth and moral responsibility. And I will compare and contrast my account to similar views, such as Haidt's Social Intuitionist Model, Virtue Ethics, John Doris's account of value-guided agency, and Angela Smith's Rational Relations View.

Section I: The Empirical Foundation Of Value-Guided Automaticity

WHAT IT MEANS TO VALUE

When I say that someone "values something," I mean that they *care about it in a non-trivial way*. To *care* about something is not simply to want it, but to have a range of emotional and behavioral dispositions to it over time.[78] For example, my caring about my younger sister involves a variety of emotions at different intervals: joy when she is doing well or I can spend time with her, sadness when she is struggling, anger at those who contribute to her struggles, appreciation for those who contribute to her well-being, pride when she succeeds, empathy when

[78] See also, Anderson (1993): To value something is to have a complex of positive attitudes toward it, governed by distinct standards for perception, emotion, deliberation, desire, and conduct. People who care about something are emotionally involved in what concerns the object of care. Parents who love their children will normally be happy when their children are successful and alarmed when they are injured. They will be alert to their needs, take their welfare seriously in their deliberations, and want to take actions that express their care" (2). ….. "My theory of value could be called a "rational attitude theory," according to which the attitudes engaged when we care about things involve not just feelings but judgment, conduct, sensitivities to qualities in what we value, and certain ways of structuring deliberation concerned with what we value." (pg. 5); Jaworska (2007, pgs. 482-483); Shoemaker (2003): emotional responses are necessary AND sufficient for caring (pg. 93); and Sripada (2015): "…caring is also associated with a rich and distinctive profile of emotional responses that are finely tuned to the fortunes of the thing that is the object of the care…. Suppose Paul cares about the plight of children disposed by war in Sudan. If a hostile United Nations resolution on Sudan is forthcoming, Paul is disposed to suite of 'signaling' emotions such as anxiety and fear that concentrate his attention on the looming threat and give it precedence over other considerations. If Sudanese children are benefited or advanced in some way, Paul is disposed to a suite of positively valenced emotions such as joy, approval, and elevation. If the fortunes of the Sudanese children are set back, Paul is susceptible to sadness, disapprobation, and despair. In this respect too, cares are quite different than desires. It is perfectly possible to desire something, but not have this rich and distinctive profile of emotional connections to the prospect of that thing being threatened, achieved, or foreclosed" (no page numbers yet, online). Though these views vary in some ways, the central thread that I find useful for my purposes is that valuing consists of a range of emotional responses and dispositions. I do not think, like Anderson, for example, that valuing will necessarily enter our conscious deliberations or that one must have conscious awareness of her values. I also do not think our cognitive capacities need to be that robust for us to be said to value. As I will discuss in more detail below, I find it possible that some animals meet the conditions of valuing described here by Shoemaker and Sripada, for example.

she fails, fear when she is endangered, relief when she is safe. I would experience great grief at her loss.[79] My caring about her also means that I am disposed to act in particular ways around her. I'm inclined to hug her when we greet, console her when she's sad, defend her when she is being criticized, make meals for her when she visits, share my snacks with her, and plan fun activities for us to do together.[80]

I care not only about my sister, but my relationship with her. If my sister and I were to have an irresolvable argument, I would still feel loss even if she continued to thrive outside of our relationship. Thus, I not only care about her well-being, but also the opportunity to connect and share with her. This point illustrates that people may care about entities other than persons. For example, one might care about a sports team, experiencing joy when the team is doing well, anger at those who pose as obstacles to the team, pride when the team succeeds, empathy when the team fails, etc. One might similarly care about a community theatre, National Park, or even a social cause like promoting animal rights or racial justice. We may even care about experiences such as traveling, challenging ourselves, spending time in nature, taking part in our city's culture and arts. In short, caring is not restricted to persons.[81]

Caring about, or valuing, someone or something also involves perceiving the subject as irreplaceable.[82] If I did lose my sister, no one would suggest I simply find another. Buying a new

[79] See also Shoemaker (2003), 92 and Sripada 2015—cares exhibit a syndrome of dispositional effects that includes motivational, commitmental, evaluative, and affective elements.
[80] Jaworska explains that caring may not necessarily involve *all* of these emotions and behaviors: "Not all of these elements must always occur in a given case of caring, but if enough of them are missing in the relevant circumstances, talking of caring is not warranted" (2007, 484).
[81] See also Frankfurt (1988), "[People] may care…about their own personal projects, about certain individuals and groups, and perhaps about various ideals..." (81); and Shoemaker (2003): "Given the connection between caring and emotions, and given the very wide and diverse range of events to which we may have emotional reactions, the range of potential objects of care is itself extremely wide and diverse" (94).
[82] Doris mentions this particular condition of valuing: "…if the object of desiring can be replaced without loss—if life can go on pretty much as it did—then that object is not an object of valuing" (2015, 28). See also Sayer "Where preferences are concerned we are generally willing to substitute something else for what we prefer. I would prefer to do x (stay with my current bank), but if I discover y (another bank) is better I might give x up for y. If I am committed to, say, my child or certain political beliefs, then they can't be sold or swapped for something else. I am

house may not fill the void of my old, beloved home. If I am saddened by the inevitable death of my sick cat, the point that I can easily get another will not console me. Even when we do seek new relationships or belongings upon losing an old one—e.g. one might search for a new partner after a breakup or death—it is mistaken to say that the new person "replaces" the old. Instead, we may find comfort in the new relationship/person as we come to terms with the loss of the relationship/person we cared about and valued.[83]

It sometimes happens, of course, that we stop caring for something or someone over time (either quickly or slowly). I will certainly stop caring about an ex in the same way or may come to lack care for him/her entirely at some point. This does not mean that the thing or person we cared for at the time was replaceable, but rather that our cares can change. When I stop caring about the thing or person, they may be replaceable to me. When our cares change like this, we may think, "perhaps I never really cared about him/her at all." This raises the interesting question of whether we can be wrong about what we care about. On the one hand, it might seem odd to say to a friend something like "even though you think you care about X, you don't" or "you say you don't care about Y, but you do." It might seem like we would be the best judgers of what we care about. However, given that my aim is to show the existence and influence of implicitly acquired and held values, I do think it is possible to be confused about what we care about—either in the present or in the past—because we do not necessarily have conscious access to all of our values and cares. Thus, I do find it coherent to say to someone "You think you really care about our friendship, but I don't think you actually do," or "You think that you don't care about equality, but you really do," or even "I thought I cared about him, but in hindsight I guess I

---

committed to certain people, ideas and causes and I can't be bought off, for they are ends in themselves, not merely means to other ends, and commitment to them, in all their irreplaceable specificity, has become part of who I am" (126).

[83] See, also, Doris (2015): "…note for now one further marker of value, non-fungibility: if the object of desiring can be replaced without loss—if like can go on pretty much as it did—then that object is not an object of value" (28).

didn't." I, in agreement with Nomy Arpaly and Angela Smith, think that we are not always aware of what we care or do not care about.

The above examples involve a deep kind of caring. People can also care about things to a lesser extent. For example, I might care about my houseplants. I feel a range of emotions depending on how they are doing—satisfaction and pride when they thrive, disappointment when they die—but my emotions are neither as varied nor intense as the emotions I experience in caring about my sister. I will not grieve the loss of my houseplants, though I will be sad if they cease to exist. And while I can easily get new houseplants, I may feel to some extent like the ones I have currently are irreplaceable. Hence, I still experience a range of emotions and dispositions, perhaps at various intervals, about my houseplants but not nearly to the same degree that I do about other things in my life. We do not value all the things we care about to the same extent.[84]

Even more minimal than this kind of caring is wanting or desiring. Caring and valuing are often described in contrast to simply wanting or desiring something.[85] Wanting and desiring involve some of the same emotions and dispositions to caring and valuing, but lacks many others. For example, if I want something, I might be happy when I gain it and sad or frustrated when it is gone. I might even be disposed to act in ways to bring it closer. But I will not feel pride, joy, anger, or grief about the object. And my dispositions may in some cases be more fleeting than in instances of valuing. For example, I may desire chocolate cake. This desiring

[84] Svavasrdottir (2014) mentions degrees of valuing, but does not go into detail (pg 109, footnote 36). Anderson (1993) suggests that there is not only a difference in degree of valuing, but also a difference in kind (10-11).

[85] See, for example, Doris 2015 "…not all desires involve values. Some desires, like my vague fancy to ask my neighbor what she paid for her new car, seem too faint to fund values, while others, like my momentary urge to call an old friend, seem too fleeting." 26; Svavasdottir 2014 "There is a difference in the kind of emotional dispositions grounded by desiring and valuing. Desire grounds dispositions such as to experience frustration, when the desired object is not attained, and to be glad, when the desired object is attained. In addition to dispositions similar to these, valuing grounds dispositions to have "deeper" emotional responses such as to regret or grieve when something adversely affects the object valued" (94-95); and Frankfurt 1988 "Thus caring about something is not to be confused with liking it or with wanting it…" (83).

entails that I feel happy when I get the cake and sad when it is gone. But my emotions are not much more complex or pervasive than this. I can also be satisfied by a replacement for the chocolate cake. If the chocolate cake is gone, but someone offers me carrot cake, I would be perfectly content.[86] Note, however, that desires and wants are not limited to physical sensations or pleasures like cake. We can desire or want things like money, material objects such as houses and cars, and social status or symbols. We can even desire or want relationships or people. John might desire a relationship with the charming and gregarious trainer at his gym, feeling happy when he gets the trainer's attention and sad when he doesn't. But wanting does not entail that John values; he might not feel anything more than happy and sad about the relationship.[87] Hence, we should think of valuing and caring as a spectrum, with deep caring—for children, parents, partners, friends, ideals, etc.—at one end of the spectrum. At the opposite end of the spectrum is not caring less, but simply desiring and wanting.[88] In between, there are things we care about to some extent, like houseplants, personal possessions, and hobbies.

A final point to make about valuing, on my account, is that valuing is distinct from thinking something is valuable.[89] For example, one could see that that something is valuable, but

[86] See also, Shoemaker (2003): "I may now desire some vanilla ice cream more than I desire some chocolate, but I may very well care about neither" (95, cf Frankfurt 1999, 157).
[87] Desiring can turn into valuing. At first John may simply want the relationship, but perhaps after time and development, John may come to value the relationship.
[88] While wanting and desiring does not necessarily entail caring/valuing, caring/valuing does entail wanting or liking. By this, I mean that caring or valuing about something or someone necessarily entails a positive affect or valence. It is not possible to care about/value something or someone, in the specific sense that I mean here, that we dislike or do not have some positive affinity for. This does not mean, however, that we cannot feel somewhat negatively about things we care about. The parent up late at night with a crying child—whom they care about deeply—can certainly feel frustrated, exhausted, angry, or even resentful. I might sometimes hate healthy eating even though I value my health. In addition to such negative feelings, on my account of caring and valuing, one must also experience some sort of positive affect—joy, satisfaction, desire to be close to, pride, etc.
[89] See also Shoemaker (2003): "…there is an important distinction between judging valuable and caring. One the one hand, to judge X valuable is not necessarily to care about X. For example, I may recognize considerable intrinsic value of a possible project or way of life, while failing in any way to be drawn to it myself…. On the other hand, to care about X is not necessarily to judge X valuable" (96, cf Frankfurt 1999, 157 and Watson 1987, 150); and Svavardottir 2014, 85. It might appear here that Shoemaker is saying that "to value" is not the same as "to care," which would undermine my description of valuing. However, I think that Shoemaker is making the same distinction

not themselves value or care about it. Someone who identifies as agnostic or atheist, for example, might be able to see the value of religion for others—it provides community, offers moral guidance, encourages compassion, for example. However, recognizing the value that religion provides does not mean that the agnostic/atheist would value religion. The distinction here is subtle: to think that something is valuable is not necessarily to value it. I might see that something has value, or is valuable to others, but not value it myself. I could recognize that money has value, but not value it myself. *I value x* is different than *I judge x valuable*; the former (on my description) involves an emotional component, whereas the latter need not. Similarly, to value something does not necessarily mean one judges it to be valuable. This second point is perhaps more contentious—it seems that one must judge something valuable to value it. However, it seems possible in at least two cases to value something without judging it valuable. First, I might value something that I explicitly reject as valuable. For example, I might value—again, in the sense I have described above, in terms of caring about something in a nontrivial way—standards of beauty, even though I explicitly reject the idea that standards of beauty have value. Despite this explicit rejection, I might find myself feeling pride when I meet certain standards of beauty and frustration or shame when I fail to meet them. Even though I may intellectually reject the value in these standards, I may nevertheless care about them in a non-trivial way. Second, as I will argue below and have mentioned above, it is possible to value something or someone without being consciously aware of it. I might value my neighbor's friendliness without even realizing it. When a friend says "wow, your neighbor is really friendly" I might shrug or fail to consciously engage with the comment. I might only notice how much I

I aim to make in this paragraph—that to judge something valuable is distinct from having particular emotional responses and engagement with it (which Shoemaker calls caring and I call valuing).

cared about the neighbor's kindness once she has moved and the new grumpy neighbor moves in.[90]

Up to this point, I have largely focused on what it means to value something, or to engage in valuing. This may be thought to be different from talking about someone's "values." When we speak of "values" we often think of things like honesty, compassion, friendship, purity, loyalty, diversity, inclusiveness, etc. But here I mean for "values" to encompass more than this. When I say "values play a central role in our moral judgment and decision-making," I mean that the things we value influence our judgments and decisions. Or, put another way, given my analysis above, our judgments and decisions are influenced by the things we care about.[91]

I turn now to empirical research on judgment and decision-making to explain exactly how this works. Specifically, I argue by analogy that such values and caring can be acquired and expressed implicitly, that is, without our awareness. I argue via analogy because it is difficult at this point to synthesize empirical research on the kind of valuing I have outlined above. This is because, first, the terms "value/s/ing" are used in a variety of ways in empirical work. For example, researchers use "value" to refer to quantified physical or emotional reward,[92] abstract

[90] It could be argued that I really do judge x valuable in these two kinds of cases, however the judgment is implicit or nonconscious (see Angela Smith's work, for example). I am open to the possibility that valuing x necessarily entails a judgment that, at some level or in some way, x is valuable if we qualify that judgments can be nonconscious and at times inaccessible. Note, however, that this still entails that one might consciously or explicitly judge x to lack value, but nevertheless end up valuing it.

[91] It might be noted that the concept "care" plays a large role in my account of valuing, so much so that one might ask whether I've properly defined "value" or whether it might be more efficient to describe my account as "care-guided automaticity." To the first question, I revisit my point that the term "value" is used in various ways which means that my analysis will apply to some discussions of value, but not others. I do not intend to suggest that one particular understanding of valuing is better (at this point, perhaps that may be important to explore in future research), but rather to carve out how I understand valuing and how this concept helps push through the stalemate in which Haidt and his interlocutors are positioned. To the second point, I find it plausible that "care-guided automaticity" better captures the account I present here. As I mentioned in a footnote earlier in this chapter, the concept of "value-guided automaticity" is inspired by Nancy Snow's "virtue-guided automaticity." Work on the more precise philosophical and empirical nature of values and caring—which I intend to pursue in the future—will help determine whether my account needs to be renamed or re-described. For now, I hope to have given a detailed enough sketch to convince the reader that, contra many responses to Haidt's work, such research is worth pursuing.

[92] See, for example, Grabenhorst and Rolls (2011); Levy and Glimcher (2012); Rangel and Clithero (2012); Barta, McQuire, and Kable (2013), and O'Doherty (2014).

ideals or guiding principles,[93] and preference.[94] Furthermore, Gregory Maio (2010) has recently argued that there are three "levels" or "layers" of valuing and Hitlin and Piliavin (2004) argue that the concept "value" has been sporadically and inconsistently used in disciplines like sociology. In short, there is little consensus in empirical research on the meaning of value/s/ing and the concept might be multifaceted. Thus, in this chapter, I draw from research on less ambiguous cognitive phenomena and suggest that if things like goals and implicit attitudes—which are more widely studied in both philosophical[95] and non-philosophical research—can be acquired, triggered, and executed implicitly, we have reason to think that values can be also. This is because I suspect that values are similar to goals and implicit attitudes insofar as they stem from and reflect, or perhaps even constitute, our evaluate structure or framework. Angela Smith (2005) argues that actions, judgments, and omissions are all morally significant insofar as they reflect our evaluative judgments/structure; I use the same thread in my analogy here. More research needs to be done on the exact nature and influence of values in the sense I have described here, but I bracket that laborious project reason to believe that some of our *value*-guided actions or judgments are caused entirely by unconscious/implicit processes. I also aim to shift the burden of proof to some degree: if cognitive processes like goals and implicit attitudes can be acquired, triggered, and executed unconsciously, why would we think that values *cannot* also be acquired, triggered, and executed unconsciously?

UNCONSCIOUS TRIGGERING AND EXECUTION OF VALUES

[93] See, for example, Maio and Olson (1998); Blankenship, Wegener, and Judd (2008); Maio et al. (2009); Torelli, Kaikati, and Carver (2009); and De Groot and Steg (2010).
[94] See, for example, Ravlin and Meglino (1987a) and (1987b).
[95] Many philosophers working on questions moral psychology and implicit bias research are familiar with researchers working on goals and implicit biases like John Bargh, Peter Gollwitzer, and Roy Baumeister.

In Chapter 2, I discussed the mechanisms of System 1 (automatic processing) and System 2 (deliberative processing) in detail. My goal in this chapter is to show that in some cases, our moral judgments and decisions are guided *purely* by System 1, meaning they cannot be grounded at any point in conscious deliberation (System 2). First, I will discuss how values can be *triggered* and *executed* automatically. In the next subsection, I will draw from work in psychology and cognitive science that gives reason to think that values can be *acquired* automatically via our implicit processes. I will then incorporate my analysis from Chapter 3 to explain how such value-guided automaticity is normatively evaluable and significant.

In "The Automated Will: Nonconscious Activation and Pursuit of Behavioral Goals" (2001), Bargh et al explain how we can activate and pursue goals through automatic processes.[96] It is commonly accepted that we can develop habits that, over time, can be *executed* automatically without conscious effort. For example, learning how to tie one's shoes requires conscious effort and deliberation. One must execute each step deliberately. However, over time, this process becomes habitual such that conscious effort is no longer required—one need not think "over, under, through" etc. Nevertheless, one must consciously activate this habit: "we are leaving, better put my shoes on." Once consciously activated, the habit can be executed without conscious thought.

Bargh et al. argue that not only can habits be executed automatically, but *triggered* automatically, that is outside of our conscious awareness and without deliberate intention (1015). They support this claim with a number of empirical studies. In these studies, Bargh et al. measure whether goals can be activated equally well both consciously and nonconsciously. Participants are given tasks such as to recall main features of a particular character, play a

[96] Here, I follow Nancy Snow (2010), who draws upon Bargh et al.'s work in her discussion of virtuous action and automaticity.

cooperative game, or complete a word search. Participants were either explicitly given goals to focus on—such as to perform well—or implicitly primed with such goals or not primed or instructed at all. Researchers found that participants who were implicitly primed with particular goals demonstrated the same outcomes as those who were explicitly instructed to aim for the given goals, and both groups followed the goals more so than the control group. Such studies lead Bargh et al. to conclude:

> …the results of the five experiments support the proposal that behavioral goals can become activated without any consciously made choice required. Once activated, these nonconscious goals operate in ways known for consciously chosen goals. They promote goal-direction action (achievement in Experiment 1, Experiment 2), they increase in strength until acted on (Experiment 3), the produce persistence at task performance in the face of obstacles (Experiment 4), and they favor resumption of disrupted tasks even in the presence of more attractive alternatives (Experiment 5). (2001, 1024).

In short, not only can goals operate automatically, but they can be *activated*/*triggered* automatically. We need not always consciously call the goal to mind—"I want to be friendlier today"—instead, our goals can be triggered and acted upon without our awareness.

Recent research on implicit attitudes illustrates the same phenomenon of automatic execution and triggering. It has been well documented in the last several decades by researchers in psychology and cognitive science that people hold both explicit and implicit attitudes, which in some instances conflict. For example, when directly asked, many people explicitly express egalitarian views about race, gender, age, ability, sexual orientation, etc.—"I believe that people are equal regardless of their sexual orientation." However, on a computer test in which participants are instructed to categorize evaluative words (e.g. "good," "evil") and people from different social groups (e.g. Black and white faces), many of the same participants who *explicitly* expressed *egalitarian* attitudes *implicitly* express *harmful stereotypes*. Specifically, people have a more difficult time, for example, associating Black faces with positively-valenced words.

Evidence from this computer test (the Implicit Association Test) is taken to show that many people hold implicit attitudes toward social groups that they are not aware of. Furthermore, research on microagressions shows that not only do many people hold such problematic implicit attitudes, but that such attitudes are often triggered and executed without our conscious awareness (Sue 2010). For example, a white couple walking home may cross the street at an earlier intersection than usual. Unbeknownst to them, the reason for this earlier crossing was because they saw a Black man on their side of the street, which triggered their implicit stereotype that Black men are dangerous, which caused them to cross the street to avoid the Black man. In this case, the couple was not aware that the implicit stereotype influenced their action, and are likely not aware that they hold the stereotype. These kinds of microagressions are an everyday example of how implicit cognitive content—in this case attitudes, associations, or stereotypes—can be *triggered* and *executed* automatically and without our conscious awareness.

Such research on the automatic triggering and execution of goals and attitudes gives good reason to think that our *values can also* be triggered and executed without conscious awareness. Given the power of implicit priming and activation illustrated in these studies, we can infer that our values can also be triggered by cues in our external environment and processed to guide particular behaviors or judgments. In fact, there is a way in which Bargh's data might not only show goal-directed automaticity, but value-directed automaticity. Instead of describing the data in terms of a goal "My goal is to cooperate," we can describe the date in terms of values "cooperation is important," or "I value cooperation." We might think that many goals are value-laden.[97] It could be argued, then, that Bargh's data establish *value-guided* automaticity in addition to *goal-directed* automaticity.[98]

[97] Nancy Snow describes this relationship between Bargh's goal-directed automaticity and values (2010, 44).
[98] This latter point would be helpful, though I do not think necessary, for the success of my argument.

The idea that our values can be implicitly triggered and executed should resonate with our everyday experiences—the concept should not feel too foreign. Imagine an example where you hear of a horrible accident or attack on the news during your morning commute. Unbeknownst to you, the exposure to such an attack makes you more affectionate to family when you arrive home at the end of the day. Your extra affection is noticed and you are asked "what has gotten into you?" You however, don't notice your extra affection and even if you did, may not be able to make the connection between the traumatic story and your increased affection. It would seem, then, that hearing the story triggered a value for you (caring about family) outside of your conscious awareness and this value was later manifested in your actions, again outside of your conscious awareness. Of course, there is a version of the story in which you *do* make the conscious connection, but certainly sometimes we don't make the connection—in other words we aren't always consciously aware of the triggering and execution of our values.[99]

In sum, there is good reason to think our values are sometimes, if not often, activated and executed automatically—in other words processed without conscious awareness.

Unconscious Acquisition of Values

While the above research indicates that goals, values, and attitudes can be unconsciously and automatically triggered and executed, it is a separate question whether goals, values, and attitudes can be unconsciously and automatically *acquired*. In fact, this idea is explicitly rejected by researchers like Bargh; he claims at various points that goals, for example, can be triggered

[99] Note that I have only given one possible explanation of the situation here. Another possible explanation is that fear is triggered by the story and I then seek to either collect or give affection to my family as a means of soothing myself. This seems less morally significant/evaluable than the scenario I described. I will not be able to say in any particular instance what exactly is driving our behaviors—in some cases it may be fear or other kinds of motivations. However, I think it is possible that in some cases, values can be implicitly triggered and executed, which is all I am trying to illustrate with this example. Given the implicit nature of the phenomena I describe, we will likely not be able to accurately identify what motivates our actions (except, perhaps, through controlled research in a lab).

and executed nonconsciously, but must be consciously acquired or developed.[100] If true, the virtue ethics-like argument discussed in Chapters 2 and 3 that automatic moral judgments get their normative force from previous deliberation (which is internalized through habit) may succeed. Thus, in this section, I provide an argument for the claim that in some cases, values are not acquired through previous deliberation. If my argument is successful, we can then turn to the question of what grounds the normativity of moral judgments and decisions guided by nonconsciously and automatically acquired, triggered, and executed values.

First, it is not clear that Bargh's claim that goals must be consciously acquired is supported by any evidence. To the contrary, research in fields like personality psychology suggests that at least some of our goals may be acquired implicitly and automatically. In a review of research on personality development, Dan McAdams and Bradley Olson (2010) describe three different layers of individual personality. In the early months of our lives, the fundamentals of unique temperaments are already identifiable as some infants are more cheerful or calm while others are anxious or distressed (519-520). These basic dispositions do not constitute personality, but there is evidence to suggest that they serve as the foundation for more complex personality traits like introversion, extroversion, agreeableness, conscientiousness, and constraint (520). For example, in a longitudinal study of 1000 children in New Zealand, statistically significant associations were found between age-3 temperaments and personality traits at age 26. Impulsive and distractible 3 year olds tended to show high levels of neuroticism and low levels of agreeableness and conscientiousness as young adults, while socially reticent and fearful 3 year olds tended to show higher levels of constraint and lower levels of extraversion (520). This is not

[100] See, for example: "…the hallmark of learned, habitual, automatic processing abilities such as skills (e.g., driving, typing) is that they were once effortful and intentional and only slowly, with considerable experience, become efficient and reflexive (see Miller at al., 1960; Newell & Rosenbloom, 1981). In the same way, the auto-motive control of behavior originates in intended, consciously chosen behavioral responses that only become habitual after frequent and consistent employment" (Bargh and Gollwitzer 1994, 73).

to suggest, however, that our personalities are genetically determined, but rather that we are born with varying temperaments that—depending on their interaction with our environment—serve as the foundation for personality traits as we age.

These traits in turn affect the kind of goals we hold in childhood and adulthood (525). For example, in a longitudinal study of 298 college students, extraverts expressed higher levels of enthusiasm for a greater number and variety of personal goals. Participants with the personality trait of "agreeableness" showed higher correlation to social and relationship goals and lower correlation to aesthetic goals. And the personality trait "openness" was positively related to aesthetic, social, and hedonistic goals and negatively related to economic and religious goals (525). While a number of our goals may be consciously acquired, as Bargh suggests, we start developing goals in early childhood and some of our goals may be the result of things like basic and social needs, personality dispositions, or experiences. For example, imagine that Alex has had a distressed temperament since early childhood. Also imagine that he did not grow up in a stable home. It would not be surprising if Alex at some point developed the goal to be in a stable-long term relationship. Furthermore, it seems just as plausible that Alex developed this goal implicitly as through conscious deliberation.[101] Similar to the sort of self-discovery discussed elsewhere in this dissertation, Alex may one day be surprised to realize, perhaps after a period of unsuccessful casual dating, that all along he has wanted more stability and security.[102]

In fact, psychologist Timothy Wilson explains that there is a whole class of motivations and goals that are acquired and held outside of our conscious awareness (2002). He cites research

[101] Just as it is possible that Alex develops the goal via deliberation, it is also possible that he develops different goals entirely given his situation. Perhaps, he implicitly aims to avoid long-term relationships, given the instability of those of his childhood, and finds himself sabotaging such relationships. My aim is not to establish definitively Alex's situation, but rather to establish that this implicit acquisition, triggering, and execution of goals is possible. The point here is similar to that I make in footnote 104.

[102] Arpaly (2003) discusses cases of this sort, where individuals may act counter to their actual preferences or goals.

(by Bargh and colleagues) on the use of sex to achieve power that showed that men who scored higher on measure of sexual aggression found a female confederate more attractive when they were primed with words that had to do with power. Wilson explains that being primed with the concept of power resulted for some men in the creation of the goal to have sex with the female confederate. When asked about the association, however, these men reported not being aware of the connection (and a number of men in the study did not express the association of power and sexual attraction at all) (35). Wilson concludes that "…people's adaptive unconscious might acquire goals of which they are completely unaware and would not act on deliberately…" (34). The idea that people can have diverging implicit and explicit motives and goals is further supported by David McClelland's psychological research in the 1980s and 1990s. McClelland and colleagues measured people's needs and motivations—for example the need for affiliation, power, or achievement—implicitly through the Thematic Apperception Test and explicitly through self-report questionnaires. Their results showed a divergence in implicitly and explicitly collected data. For example, a participant might *explicitly* report a high need for affiliation, but show on *implicit* measures a low need for affiliation (Wilson 2002, 82-83). In turn, people may have diverging implicit and explicit motives and goals. Wilson claims that implicit motives and goals are often derived from needs that are acquired in childhood that now operate automatically and nonconsciously (83). Additionally, Wilson explains that *implicit* motives better predict people's behaviors than do their *explicitly stated* motives (83). Given this research, Bargh's claim that goals must be acquired consciously is unsubstantiated. There appears to be evidence that some of our goals *are* acquired outside of our conscious awareness.

One final point about goal acquisition is that given the methodology used in many studies on goal triggering and execution, it is not clear whether researchers like Bargh are observing *only*

goal triggering and execution *or* goal acquisition, triggering, and execution. For example, in one study, researchers briefly exposed participants to a short script that implied the goal of earning money. In a later task, researchers observed whether participants would strive to finish the task more quickly given the opportunity to participate in a lottery. Researchers found that participants primed with the short script did in fact aim to finish the task more quickly, though without conscious awareness. Psychologists Dijksterhuis and Aarts describe this phenomenon as "goal contagion," meaning goal inferences "occur spontaneously, without conscious intentions and awareness" (2010, 473). However, it is not clear if the phenomenon witnessed is merely the implicit *activation* of a goal—to earn money—or the implicit acquisition *and* activation. To argue that only activation is occurring is to presume that the participants already had the goal of earning money. This assumption needs further analysis and support. Dijksterhuis and Aarts also describe a study where participants exposed to a Mac computer led to the goal to be creative in comparison to exposure to an IBM computer (473). Given that the goal of being creative is perhaps less common than the goal of earning money, I find it plausible, if not likely, that participants are not having an existing goal triggered, but rather being primed to acquire, or take on, that goal. If true, there would be evidence, contra Bargh, that goals are sometimes not only triggered and executed implicitly, but also implicitly acquired.

In addition to this research on goal acquisition, there exists much research in psychology and cognitive science about other kinds of cognitive content, attitudes, knowledge, skills, emotions, and concerns that we develop or acquire implicitly without conscious awareness. In the remainder of this section, I will discuss research on implicit learning, implicit biases, attachment skills, falling in love, and developing concern for others.

In "Implicit Learning" (2003), psychologists Peter Frensch and Dennis Runger describe several key studies that suggest people have the ability to acquire information and knowledge without conscious awareness. In one study, participants are asked to memorize a set of letter strings, such as "XXRTRXV" and "QQWMWQP" (Frensch and Runger, 13). These letter strings follow particular patterns and rules, though the participants are not made aware of these patterns. After the memorization phase, participants are told that the previous set of letter strings followed a pattern and are asked to complete a new set of letter strings. Participants are often able to complete the new letter strings, indicating that they learned the relevant pattern, however, they are unable to articulate the pattern when prompted. This suggests that participants did learn the pattern, but that such learning happened implicitly, or outside of their conscious awareness. A similar study measured participants' ability to learn spatial patterns with similar results. These classic studies show that implicit learning is possible, making plausible the idea that some of our values are not only triggered and executed without conscious awareness, but also *acquired* or learned without conscious awareness. Notice here that I am making a different point than that made in previous chapters. The point is not simply that we are implicitly influenced by external factors—priming, environmental stimuli, etc.—but that we can acquire knowledge that may persist over some period of time without being consciously aware of it.[103]

Research on implicit attitudes and biases also suggest that they are developed implicitly rather than explicitly at a young age and are sustained through adulthood even as our explicit attitudes change. In "The Development of Implicit Attitudes," (2006) Baron and Banaji explain that 6 year olds of various racial groups show implicit and explicit biases that favor dominant groups, like white people. Hailey and Olson (2013) claim that these biases are also found in

[103] The classic Iowa Gambling Test may be another useful example here, where participants seem to unconsciously identify riskier decks before being able to consciously identify them.

children at ages 4-5 (461). At ages 10, children still manifest the implicit biases, though the explicit attitudes begin to fade. By adulthood, explicit attitudes in many people have completely changed to be more egalitarian, however the implicit biases still remain as measured by the Implicit Association Test. These implicit biases, which are developed without conscious awareness or deliberation, can continue to influence our judgments and decisions.

Another kind of emotional and cognitive experience that develops without conscious deliberation, but nevertheless is normatively significant is falling in love. Falling in love involves a swath of implicit processing of implicit information, such that we often experience falling in love as something that happens to us rather than something we deliberate about or choose. We often find as we share experiences with someone over time that we come to love them. This happens as we implicitly and explicitly gather information about another person. But the process of falling in love seems in large part to occur in System 1—meaning it is intuitively driven, happens outside of our conscious awareness, and cannot be articulated. We might be able to identify features that attract us to certain partners or people "I was attracted to her because she was very kind," "she was adventurous," "he was caring," etc. but we often cannot explain why we fall in or out of love with each other. It is not uncommon to hear "I can't help but love her," or "I just don't love him anymore. I don't know why." It might also be possible to fall in love with someone without being aware of the fact that you've fallen in love. You may come to the realization that you love someone and perhaps have loved them for a long time. In a sense, you have come to value them in a particular kind of way—you have developed a new value—your partner or your relationship—via implicit processing.[104]

[104] See Iris Murdoch, for example, on the implicit acquisition of moral values and commitments: "…if we consider what the work of attention is like, how continuously it goes on, and how imperceptibly it builds up structures of value round about us, we shall not be surprised that at crucial moments of choice most of the business of choosing is already over. This does not imply that we are not free, certainly not. But it implies that the exercise of our freedom is

Caring and valuing oftentimes have a similar kind of phenomenological experience to falling in love. We often feel like we cannot help but value the things we do and could not stop valuing or start valuing by sheer will. David Shoemaker (2003) describes both of these points in detail. First, he discusses Frankfurt's 1988 example of Martin Luther taking a stand against the church, supposedly declaring "Here I stand, I could do no other" (Shoemaker 2003, 104; cf Frankfurt 1988, 80-94). This example is used to illustrate the phenomenon of being compelled to act by our cares. I think this is in fact a useful way to describe the cases I have discussed in this dissertation. When Wesley Autrey dodges the question *why did you do what you did?*, I suspect it is in part because he has the sensation "I could have done no other." The same may be said for those who helped Jewish refugees and even Huck Finn (though Huck sees this as a failure of will).[105] When a friend thanks me for helping out in a hard time, I say, "please, *of course* I will help. It is not even a question." Like falling in love, acting out of care is in some cases not something we choose or deliberate about, but almost something that happens to us.

Relatedly, Shoemaker claims, departing from Frankfurt, that "none of one's current cares are under one's voluntary control at the current time" (2003, 105).[106] By this, Shoemaker means that we cannot will to start or stop caring about something through conscious deliberation in any given moment. My roommate broke my favorite glass. I may know that I shouldn't be as upset as I am, but nevertheless cannot stop feeling the way I do. I care about the glass and her negligence as hard as I might try to stop caring. Similarly, my physician might tell me that I need to be exercising more to lower my cholesterol. However, much as I try, I just can't seem to muster the

---

a small piecemeal business which goes on all the time and not a grandiose leaping about unimpeded at important matters. The moral life, on this view, is something that goes on continually, not something that is switched off in between the occurrence of explicit moral choices. What happens in between such choices is indeed what is crucial (Sayer 2011, 97; cf Murdoch 1970, 36).

[105] Shoemaker argues that acting from volitional necessity is not a failure of agency, but rather an expression of it (2003, 105).

[106] Contra Frankfurt: "When a person cares about something, it may be entirely up to him both that he cares about it and that he cares about it as much as he does" (1988, 85).

care about it. Shoemaker does acknowledge that we *can* change our cares over time—I am finally able to let go of the broken glass and find myself less upset when a broken plate appears. However, Shoemaker says that this change of caring is caused by some other care I have at the time, perhaps wanting to have a better relationship with my roommate, wanting to be less emotionally reactive to accidents, wanting to be less attached to objects, etc. Hence, changes in our cares will often be brought about by existing cares (note that this can happen explicitly through conscious deliberation or implicitly), the point being for my purposes that there do seem to be many cases in which we do not choose our values, or what we care about, but instead can acquire them more passively.

I introduce these various examples not only to show that we can implicitly acquire values and concerns, but also to show that our System 1 process can do much more than Haidt implies. Haidt makes it sound as if our automatic processes are merely reflexive, like a knee-jerk patellar reflex—with simply inputs and outputs. Input disgust and you will get a judgment that consensual incest is morally wrong. As I stated in Chapter 2, in these terms, theorists are not necessarily wrong to be skeptical of the plausibility of Haidt's account. However, Haidt has arguably misrepresented the intelligence and potential of automatic processing. Automatic processing can help us acquire knowledge without conscious awareness or deliberation, can help us develop attachment to others, and can acquire and execute important values and emotional states—not simply disgust, sacredness, and purity, but also love, care, appreciation, etc. These automatic processes can be responsive to moral reasons and adaptive to environmental and internal cues.[107]

[107] See also, Nancy Snow: "Nonconsciously activated goal-directed behaviors are not reflex reactions to stimuli, but are intelligent, flexible responses to unfolding situation cues and display many of the same qualities as consciously chosen actions" (2010, 43).

One might still press here that the value-guided automaticity I have described is just as thin or reflexive as the kind of automaticity Haidt has described: my argument that implicit, System 1 processes engage with moral values does not necessarily establish that the automatic judgments or actions are reasons-responsive, but simply that moral values can be the input of the system. Furthermore, if those moral values are themselves implicitly acquired, it might seem as if actions that are entirely guided by implicit processing, even when engaging with values, are merely trained or determined.

However, that System 1 engages with moral values makes it more responsive than Haidt allows. As I have described System 1, it responds to evidence, to the things we care about. A non-responsive system would produce the same output regardless of the things we cared about, desired, or wanted. For example, Neil Levy discusses the "alien hand" phenomenon (2014). This is a disorder that occurs in people whose two sides of the brain do not properly communicate with each other. This disorder causes people's hands to move, grab, hit, steal, etc. without their awareness or control. Levy explains why these actions are not said to properly reflect the agent:

> "These representations that guide this behavior are 'thin': they are responsive only to a very narrow set of features. Correlatively, they cause behaviors in ways that are inflexible, since the behavior is not modulated by features outside this narrow band. Because the representations that guide the alien hand are thin, there is a strong case for refusing to identify the movements they cause with the agent. An agents's [sic] propositional attitudes should not be identified with thin representation; rather they are built up out of sets of representations…. Behaviors caused by thin representations are not expressions of our evaluative stance, and therefore fail to express our good or ill will. (2014, 24-25).

I have argued above that in at least some cases, System 1 engages with moral values, which constitutes a more robust process than something like the process driving the alien hand phenomenon. Caring, even when it happens automatically, without being consciously chosen, is more substantive than the patellar reflex. I think that the impulsive action of hugging a distressed

stranger on the subway, even if done automatically and without conscious choice (perhaps by someone quite surprised by their action later upon reflection) is more robust, complex, and responsive than a sneeze, sweating from nervousness, or a shove when someone jumps out at you. As I will go on to argue in this chapter and the next, the triggering and execution of our moral values, even when unconsciously acquired, have substantive normative standing, especially when compared to phenomena like alien hand syndrome.[108] Also, contra Haidt, my analysis shows that are moral judgments and behaviors are not merely evolutionarily or instinctually driven, but in some cases driven by substantive moral content.

My goal in this section has been to establish the following:

1. That it makes sense empirically to say we have values—that is, we genuinely care about things
2. That automatic processes sometimes draw from these values to produce moral judgments, decisions, and actions (activation and execution)
3. That automatic processes can instill these values and cares (acquisition)
4. That the processes thus engage with and utilize moral reasons, and are more than atavistic, crude reflexes.

Given these points, it is now appropriate to ask whether such value-guided automaticity can be normatively evaluable. I turn to this question in the next section. Recall that my ultimate goal is to show that the fact that automaticity guides some if not much of our moral judgment and decision-making does not undermine our agency or status as morally responsible beings.

Section II: The Normative Standing of Unconsciously Acquired Values

In Chapter 3, I argued that good moral judgment and decision-making need not necessarily be grounded in deliberation. Instead, what we care about when it comes to evaluating moral judgments and decisions is whether they are founded upon the right moral reasons. Why

[108] More empirical work would help strengthen my point here, so I intend to pursue this point in future research. In the meantime, I hope my comments satisfy the reader enough to continue with my analysis.

did the CEO donate to Oxfam? Because she wanted to help. Why did she want to help? Because she cares about the suffering and well-being of others. This seems like a good moral reason for donating to Oxfam. "Because she wanted recognition" is not a good moral reason. In the previous section, I have argued that automatic intuitions can be derived from implicitly developed and held values or concerns. Can such values and concerns constitute the "right moral reasons" for judgments and decisions?

First, let me revisit Markovits's idea that what matters when evaluating an agent's motivating reasons is her noninstrumental reason—which is a reason for which there is no alternative motivation. "They needed help" is a noninstrumental motivating reason. It may not be a *foundational* reason—we could still press one about why "they needed help" is significant, perhaps revealing deeper reasons like "we should minimize suffering when possible—but it is noninstrumental insofar as it does not serve some other motivating reason. (Markovits 2010, 230). The first question that needs to be answered for my analysis is whether values and cares can serve as noninstrumental reasons. I answer "yes, they can." As I explained in Chapter 3, a noninstrumental reason is one that does not entail any other motivating reason. For example, if we ask what was the reason Jack volunteered at the hospital, he might respond that it made him feel good. It would be coherent to ask why did it make him feel good, to which he might further respond, because he liked making people feel better. We could further ask why he liked making people feel better, to which he could respond that he cared about their well-being. If we further pressed about why he cares about their well-being, Jack would likely reasonably respond "I just do," signaling that his caring is the noninstrumental motivating reason for his action. There might be more fundamental reasons for Jack's caring, but recall that Markovits says we *only* need to follow the chain of reasons down the noninstrumental reason; demanding fundamental

reasons is too stringent. In this case, Jack's caring, or valuing, does serve as a noninstrumental reason for his judgments and actions.[109] In fact, many of our noninstrumental reasons may involve caring and valuing, as suggested by Chandra Sripada (2015a):

> When we look closely at the structure of a person's economy of motivational attitudes, we find that many of these attitudes are arranged in an *instrumental* hierarchy; that is, the person desires to do X only in order to fulfill her desire to do Y, which in turn is in the service of fulfilling desire Z. Consider Katya, who wants to get on the bus. She does this only because she wants to get to class, and this too is done in the service of further desires: fulfilling the organic chemistry requirement, getting her medical degree, becoming a competent physician. When we trace sequences such as these to see where they lead, we often encounter at their very root a distinct class of conative states: *cares*. Katya wants to be a competent physician because she *cares* about helping those who are in need—she wants to relieve their suffering. (no page numbers yet, italics original, published online August 2015).

I wire my sister money without thinking twice. Why? Because I want to help her. Why? Because I care about her. That I care for her would be, on Markovits's terms, the noninstrumental fundamental reason for my action; I do not care for her for some other motivating reason—for example, that I hope she'll take care of me in my old age or promised our parents I would—my care is intrinsic and I cannot feel otherwise.

Thus, caring can constitute noninstrumental motivating reasons. Furthermore, caring or valuing can constitute *good* noninstrumental reasons of the kind Markovits says are required for moral worth and praise. On the one hand, this appears very intuitive and in need of little argument. Doing things out of care seems like exactly the right reason to act, in direct contrast to

[109] While I use Markovits's language here, it should also be noted that my view also reflects what are called "real-self" or "deep-self" views of moral responsibility. Sripada (2015c) describes a real/deep-self thesis as: "Morally responsible action is action that expresses not from any old desire, but rather the attitudes that constitute the person's self" (3). Sripada offers a nice discussion of the history and variety of theorists writing in the real/deep-self tradition, such as Aristotle, Hume, Dewey, Gary Watson, Harry Frankfurt, David Shoemaker, and Agnieszka Jaworska. (2015c, 3-5). Real-self accounts of moral responsibility are often contrasted to accounts of reasons-responsiveness. However, Sripada argues that deep-self and reasons-responsiveness accounts of moral responsibility entail surprising significant overlap, insofar as both emphasize fundamental connections to the self (2015c, 1). Hence, though I take my account to continue on in the tradition of deep-self views, I do not expand upon the connection at this particular point.

doing things out of greed, fear, or apathy (all typically bad moral reasons for acting). However, caring or valuing in and of itself cannot be a good moral reason for acting because people can care about or value the wrong things. Take, for example, someone who cares a lot about his appearance (Robin). Given this value, Robin spends his mornings working on his appearance rather than helping his partner get the kids ready for school. In this case, it seems like caring (about his appearance) is *not* a good moral reason for failing to assist with childcare. Even if Robin were to do a morally good thing like donate his old clothes, if he did so because he wanted to maintain his attractive appearance (versus wanting to helping people in need), he fails to act for a *good* moral reason. Hence, simply caring or valuing cannot constitute *good* moral reason for acting, one must care about or value the *right* things.

I do not attempt to answer the question here about what we ought to value. My aim is not to establish what makes an action right or wrong, but instead to show that moral responsibility is anchored in what we value, where the details of what ought to be valued are determined by ethical theorists, not moral psychologists. Let me nevertheless mention that I think my account is amenable to ethical theories such as deontology, utilitarianism, virtue ethics, care ethics, and rights-based approaches, for example. It doesn't matter on my account whether one thinks that we ought to care about human dignity or sentient suffering, I have only argued that such cares can be acquired, triggered, and executed automatically outside of conscious awareness.[110] While I have, following many theorists in this literature, made my own assumptions here about what is and isn't appropriate to value (e.g. I assume valuing family is good, fashion less so) and what

[110] Some ethical theorist, perhaps most notably in the traditions of deontology and virtue ethics, might nevertheless have a problem with this claim that I am making. On some interpretations of these accounts, moral action requires not only that we act for the right reasons (respecting human dignity, expressing virtue), but also that we consciously recognize these right reasons. I think it is an open question, however, whether conscious deliberation is necessary for moral responsibility on these accounts. See Markovits (2010) for an articulation of why conscious deliberation is not necessary for Kantian moral responsibility and Snow (2010) for a reconciliation of virtuous action and automaticity.

constitutes good and bad reasons (e.g. preventing suffering is a good reason, making profit for one self may not be), I think many of the cases, values, and reasons I have discussed are fairly uncontroversial. In cases where it is less clear what constitutes good or bad moral reasons, I leave it to the reader to pick their favorite moral theory and apply it. This is how many other theorists approach the issue as well. In the literature on moral responsibility, there is much said about what makes us morally responsible for our actions, but little said about what makes us praise- or blameworthy for our actions. Angela Smith describes the distinction:

> I interpret the fundamental question of responsibility as a question about the conditions of moral attributability, that is to say, the conditions under which something can be attributed to a person in the way that is required in order for it to be the basis for moral appraisal of that person. To say that a person is responsible for something, in this sense, is only to say that she is *open* to moral appraisal on account of it (where nothing is implied about what that appraisal, if any, should be). (Smith 2005, 238; italics original).

Like Smith, I am not outlining conditions of praise and blame here, as again that would require a theory of what makes an action right or wrong, but rather arguing that conscious deliberation is not necessary condition of moral responsibility.

Though I take my account to be consonant with different moral theories, I will say a bit about why I think care ethics in particular helps illuminate the questions engaged here, in contrast to other ethical theories. First, care ethics emphasizes the centrality of care in our moral judgment and decision-making (Held 2006, 10). This emphasis makes my account seem prima facie plausible: valuing and caring, even when implicit, have moral significance and can—perhaps *should*—serve as the foundation of evaluations of moral responsibility. Another feature of care ethics that complements my project is the rejection of the idealized *rational* moral agent, replaced instead with a picture of a moral agent who not only experiences emotions but is in part

morally guided by those emotions (Held, 10).[111] I have argued here for cases where individuals are morally praiseworthy because they act upon their values/cares, even though they have not deliberated about such values/cares. They acted because "it felt right" or they were compelled to do so. The Care Ethicist's point about the role of emotions in attuning and guiding us morally supports the kind of automatic, responsive model I have presented here. Third, Care Ethics rejects the idea that moral judgment and decision-making must be impartial (Held, 10). The idea that we are deeply embedded in and constituted by our relationships, which sometimes ought to take priority in our moral judgment and decision-making, complements my claim that valuing persons, relationships, communities, etc. can serve as a good noninstrumental reason for one's action—that I care for my sister is a morally appropriate, if not paradigmatic, reason for acting. There is more to say about what exactly my account draws from and adds to care ethics, yet for now I only hope to show that care ethics is a useful framework for thinking about the questions raised in my project and lends some credibly to the account I have presented.

Thus far in this section, I have established that a) valuing/caring can constitute noninstrumental reasons for acting and b) valuing/caring can constitute *good* moral reason for acting. To rephrase the point in Markovits's terms: an action is morally worthy whenever the noninstrumental reasons for which it is performed (which I have argued include valuing or caring about something or someone) coincide with the noninstrumental reasons that morally justify its performance (which again can include valuing or caring about something or someone).

The final point to reiterate about my account here is that in many cases our values are acquired, triggered, and executed outside of our conscious awareness. Thus, the model of good

[111] Held writes: "Second, in the epistemological process of trying to understand what morality would recommend and what it would be best for us to do and to be, the ethics of care values emotion rather than rejects it. Not all emotion is valued, of course, but in contrast with the dominant rationalist approaches, such emotions as sympathy, empathy, sensitivity, and responsiveness are seen as the kind of moral emotions that need to be cultivated not only to help in the implementation of the dictates of reason, but to better ascertain what morality recommends" (10).

moral judgment and decision-making that I have outlined above can occur free from the influence of conscious deliberation. I posit that is what happens for Huckleberry Finn, the Holocaust rescuers, and perhaps Wesley Autrey. When faced with moral situations, they act upon implicitly developed and held values. Huck values Jim as a friend, though has yet to realize it. People who hid Jewish refugees value others so much that they are willing to risk their own safety and their family members' lives. The same may be said for Autrey. Recall that Angela Smith explains that we sometimes *discover* that we care about things when confronted.[112] This account would help explain why such individuals are not able to properly explain their actions. Contrary to Haidt's suggestion, they are not failing to act for moral reasons, but instead are acting upon implicitly developed and held values which they have only learned they have through the situation at hand. Following the analysis I provided in Chapter 3, then, these judgment and decisions guided by implicitly developed values would meet the conditions of moral responsibility, because these values would serve as either the right or wrong relevant motivating reasons.

Note that I do not intend to imply here that we are responsible for *all* of our nonconscious decisions and actions. Many of our unconscious states and processes *are* thin, inflexible, reactive—in short morally insignificant. I might crinkle my nose at deviled eggs because I got sick a few months ago right after eating some. This is purely an association—I do not crinkle my nose because I do not value deviled eggs or even because I value my health (I might have the same automatic response even if I don't care about my health all that much). It is purely an associative reaction. My aim here has been to show how some of our nonconsciously produced

[112] See also Sripada (2015a) for more on the idea that we can care about something without knowing it: "…it bears emphasis that, on the care-based view of the deep self, a person can always be mistaken about the contents of her deep self. What makes a mental state a deep attitude is its exhibiting the relevant suite of functional role properties. It does not matter what the person thinks about the state, whether she regards it favorably, whether she consciously identifies with the state, or whether she recognizes the attitude as deep" (no page numbers yet, close to footnote 16).

judgments and decisions, however, are more robust than these simple kind of associations.[113] This is the main difference between my account and Haidt's. As I discussed in Chapter 2, Haidt describes the automatic system as somewhat reflexive, lazy, and almost purely responsive to social rather than moral reasons. I have argued here, in contrast, that our automatic systems are emotionally intelligent and engage with moral reasons regularly, focusing particularly on cases where automaticity engages with what we value. Let me turn now to explain how this account differs from Virtue Ethics, Angela Smith's Rational Relations View, John Doris's value-guided agency.

It may be suggested that I have not outlined a new account of moral judgment and decision-making here, but rather described in different words a virtue ethics account. Many virtue ethics accounts also hold that much of our moral judgment and decision-making happens automatically, often without conscious thought. However, my account differs from virtue ethics in two important ways. First, according to virtue ethics, one must consciously deliberate about morality before good moral judgment and decision-making can become internalized and automatic. As Bargh says about the triggering and execution of goals, virtues must be consciously cultivated and may at later points be implicitly triggered or executed. Here, I have attempted to show that our values need not be consciously cultivated for them to serve as a foundation of our moral judgments and decisions. Second, to do the right thing on a virtue ethics account, one must develop *virtues*. My account differs insofar as I have focused here on values, which I take to be much less robust than virtues. In a way, then, my account is less demanding

[113] See Sripada (2015a) for a similar point, where he distinguishes between implicit attitudes (which he sees are associative and reactive) and non-conscious attitudes that are based in care (45-47). See also Shoemaker 2003: "We all do a variety of things each day that seem to bear no dependence relation at all to our cares; for example, we get out of bed, we scratch itches, we reach for the milk, we change the TV channels, and so on. These are all intentional, motivated actions, explained (rendered intelligible to certain of our desires), one might say, without any necessary reference to things we regard as important, things whose changing fortunes tug on our emotional tethers" (97).

than a virtue ethics account. Actions need not stem from full-fledged virtues, but rather from values, which I posit are more common than virtues. Hence, the threshold for morally worthy judgment and action is easier to meet in my account than virtue ethics.

In Chapter 3, I discussed Angela Smith's Rational Relations View, according to which one is morally responsible for attitudes and actions, even those that we do not explicitly choose. Smith argues that we are responsible for involuntary attitudes and actions (or omissions) because they stem from our evaluative judgments, meaning they reflect what we care about or regard as important or significant. This sounds similar, if not identical, to the value-guided account I have described above, where we are responsible for moral judgments and decisions that stem from our values (things we care about), even if we did not consciously acquire, trigger, or execute those values. However, my account is different from Smith's in three ways: first, Smith is focusing on whether *choice* is a necessary condition of moral responsibility. She positions herself in contrast to the volitional view of moral responsibility, according to which choice, decision, or susceptibility to voluntary control is a necessary condition of responsibility (Smith 2005, 238). While my project is similar, I focus on whether conscious deliberation is a necessary condition of moral responsibility, thus engaging different literature and a different focus.

Second, Smith's account is anchored in "evaluative judgments" while mine is anchored in values. While these two concepts may seem the same on the face of it, Smith's "evaluative judgments" have more cognitive content than I think "values" do. Smith does explain that evaluative judgments need not be propositional nor consciously held (2005, 251), describing these judgments instead as "tendencies to regard certain things as having evaluative significance" (251). These judgments, taken together, make up our evaluative framework (251). Smith indicates either that only humans can hold such evaluative judgments, or that only humans can

be answerable for their evaluative judgments (256). Given this, she explains, nonhuman animals are not responsible for their attitudes according to the rational relations view (256).[114] In my account, however, nonhuman animals can hold values—insofar as they can care about things and have complex emotional experiences and dispositions—and thus they may be morally responsible for some of their actions. However, I do not think it is plausible for humans to assign such responsibility or hold animals morally responsible given our inability to communicate with and understand animals' motivations and mental experiences. Nevertheless, I find it plausible that some animals (perhaps advanced primates, dolphins and whales, elephants, etc.) may display moral action that is founded upon their valuing—for example grieving a fellow being's death.[115] And while humans cannot assign or hold responsible, it seems that individuals within certain species do hold each other responsible and engage in practices and praise and blame. In cases in which these are merely evolutionary/instinctual practices, I would not say they are genuinely

[114] Jaworska (2007) and Shoemaker (2003) also appear to reject the idea that nonhuman animals can express agency: Shoemaker, following Watson 1987, claims that functioning capacities such as the ability to envisage or see the significance of certain alternatives, reflect on oneself and the origins of one's motivations, comprehend or respond to relevant theoretical "are surely a precondition of genuine agency; if they are not in place, on is simply not an agent—or at most one is a severely damaged agent…If one's *ordinary background cognitive mechanisms*—for critical self-reflection and evaluation, for having one's cares and subsequent motives be responsive to such reflection, and for having certain of one's other cares tapped into in relevant circumstances—are incapacitated, then the conditions necessary for agency, and thus freedom, are absent" (2003, 117; emphasis mine). See also Arpaly (2003): "April, a child, discovers that the family dog destroyed her dinosaur-shaped toy. She becomes angry; "But it's *my favorite dinosaur*!" she screams. We may well imagine a parent explaining to her, "He's only a dog, darling. He does not understand that it's your favorite dinosaur." The dog does not understand *mine, favorite,* or *dinosaur*, not even in the murky, visceral way a small child does. Similarly, the dog's mind presumably cannot grasp—nor can it tracks, the way even unsophisticated people can—such things as increasing utility, respecting persons, or even friendship. As Hobbes hints, even if some protoversions of these notions exist in the animal's mind, these are not concepts that it can sophisticatedly apply to humans. Thus, even if this animal can act for reasons, to some extent, it cannot respond to *moral reasons*, even though it may occasionally come close" (146).

[115] Sripada 2015a also suggests that his account better includes "agents at the margins" who may not have the same reflective capacities as cognitively-typical human adults, in contrast to more reflectively demanding accounts "Agentially demanding approaches to self-expression additionally face a third family of problems: They have difficulty accounting for certain 'agents at the margins', for example young children and people with certain disability, who appear to be able to express their selves in action" (above footnote 20). Sripada does not explicitly mention intelligent animals, but it seems from his examples that he might: "Consider a young child's comforting a crying parent; a partially senile woman's enjoying helping to prepare a meal for her family; an autistic man's spending the whole month continuously reading about dinosaurs. These agents appear to have things that they genuinely care about, and their respective actions seem to be expressive of their cares" (insert page number).

moral. But given some species' emotional, cognitive, and social capacities, it seems plausible that they, in at least some instances, can be genuine moral agents.

Finally, and relatedly, Smith's account seems to entail an "articulation requirement" of the sort I have discussed in previous chapters. In other words, it appears that Smith holds that an individual ought to be able to articulate her reasons for holding a particular attitude or making a particular decision. Smith says, for example: "…once one realizes that one holds a certain evaluative judgment, it is open to one to determine whether one has adequate justification for that judgment and to modify it or give it up if such a justification cannot be provided" (252). She also writes that "…moral criticism addresses a person qua rational agent and asks her to acknowledge and to defend of disavow the judgments implicit in her responses to the world around her" (256). In some cases, people will be able to either defend or disavow their evaluative judgments. However, a particular kind of case I have been interested in is where the individual is neither able to defend nor chooses to disavow their judgment or behavior: "I don't know, it's just wrong." I have argued that one's inability to defend a particular judgment or decision does not necessarily mean that the individual does not hold good reasons for that judgment/decision. Instead, it is possible that these reasons are operating outside of one's conscious awareness. It is not clear, from Smith's current writings, whether she would be open to the possibility that an individual would not be able to defend an attitude/judgment/action, but would also be justified in continuing to hold it. Here, then, my account differs from Smith's insofar as I explicitly reject the articulation requirement in justifying one's judgments and behaviors.

Smith's project is framed slightly differently and engages with a different set of literature. However, John Doris's valuation account of moral judgment and decision-making, like my account, is framed as a response to reflectivism, taking many of the same questions and research

as its starting point. His thesis also echoes mine: Doris writes that behavior is self-directed (i.e. agential, meaning we are responsible for it) when it expresses one's values (Doris 2015, 25). He phrases this in several other ways:

> "Attribution of agency and responsibility may be warranted when a patterns of cognition, rationalization, and behavior emerges, and that pattern is best explained as involving the expression of some value." (164).
> "…when someone's deed manifests their values, it makes good sense to direct anger or admiration their way" (165).

As in my account, there are two important concepts to cash out here: what is means to value and what it means to say that a behavior expresses our values. Doris identifies a number of features of valuing—concluding that "values are associated with desires that exhibit some degree of strength, duration, ultimacy, and non-fungibility, while playing a determinative-justificatory role in planning" (28). In this definition, Doris aims to distinguish values from desires, where *mere* desires are states that are more often faint and fleeting (26), instrumental (26), and non-determinative or justificatory—that I desired cake was not a good (justificatory) reason for stealing my niece's slice (27). As I noted above, when we merely desire something (versus valuing) we are often satisfied with a replacement/substitute—my desire for cupcakes can easily be satisfied by brownies if the cupcakes are gone (28).[116] Finally, Doris explains that we need not be aware of our values to have them nor be aware of their role in our moral judgment and decision-making for our actions to truly express them: "…people may have desires, values, and plans that they are quite unaware of, and so their behavior may express their values without their knowing that it does so" (27-28). Upon describing the nature of valuing, Doris concludes: "A

[116] Note that my not being satisfied with brownies, and insisting on cupcakes, would not indicate that I *value* the cupcakes, but rather that my desire is specific enough that brownies may not suffice. However, if I went to grab the cupcake to the left, but was preempted by someone else and thus forced to grab the cupcake on the right, my desire would still be satisfied. Thus, the non-fungibility can apply to individuals or types.

behavior expresses a value, we can say, when that behavior is guided by a value-relevant goal." (26)

While Doris's account sounds remarkably similar to mine, there are three important differences. First, like Smith, Doris's account has a higher cognitive standard than mine as indicated by his exclusion of nonhuman animals in the realm of responsibility. He writes: "I'm not tempted to see my rowdy puppy as self-directed because I'm not tempted to see his behavior as an expression of his values (I don't think he *has* values)" (25) and "Normal human adults have cognitive and psychological capacities that are necessary for the exercise of agency, which other critters lack, and normal healthy children will eventually have such capacities, while other critters never will" (39).[117] As I explained above, my understanding of valuing makes possible that some animals will in fact value and thus may be said to be morally responsible for certain actions.

Second, and more importantly, Doris seems to conflate moral and nonmoral values in his account, making too broad his claims that:

> "Attribution of agency and responsibility may be warranted when a patterns of cognition, rationalization, and behavior emerges, and that pattern is best explained as involving the expression of some value." (164).
> "…when someone's deed manifests their values, it makes good sense to direct anger or admiration their way" (165)

Examples of values Doris discusses include chic and chastity (26), friendship and property rights (160), and carnal pleasure and sexual fidelity (162). Only in some of these cases would we hold someone morally responsible for their value. For example, when one's action expresses valuing of friendship (as Doris also thinks happens in Huck Finn's case), we may say he is morally

[117] This is a bit confusing because Doris writes in the following paragraph: "…nothing about assuming that adult humans are the paradigmatic agents…entail[s] denying that some of the known non-human organisms—smart animals such as dolphins, apes, and elephants—may, sometimes, to some extent, exercise agency" (39-40). More explanation here on his part would be helpful.

responsible for that action. Huck is morally responsible, and in turn praiseworthy, for protecting Jim. However, it is not clear that one's valuing fashion or carnal pleasure has any significance for the moral status of their judgments or actions. It is true that Robin's deciding to wear the striped sweater to the party is an expression of his valuing style, but his behavior ought not elicit a moral evaluation simply because it is an expression of his value. Robin's decision is self-directed in a nonmoral sense, he is the agent of his action, but "that he values fashion" does not seem to warrant, as Doris suggests in the above quote, that we "direct anger or admiration" his way. Likewise, valuing carnal pleasure in and of itself does not seem to indicate anything about our moral responsibility, as with valuing a tidy home, excellence at one's job, or being social. These values only have significance if they conflict with or take priority over morally relevant values. For example, when one betrays a relationship by engaging in sexual intimacy outside of it, the wrongdoing is not that one values "carnal pleasure" but that one does not properly value their relationship. Or when one snaps at a spouse for leaving a mess in the kitchen, the wrongdoing is not that one values a tidy home, but that does not properly value kindness or patience. In other contexts, where carnal pleasure does not take priority over a committed relationship—but perhaps reinforces it—the expression of that value is not morally relevant. Hence, it is not the case that the expression of any value meets the conditions of responsibility. Some values are not relevant to responsibility (specifically nonmoral values). Also, in some cases, it is not the expression of the value that makes one morally responsible, but the failure to express the morally relevant value in this case (e.g. patience, friendship) that serves as the content of moral evaluation.[118] In my account, I focus specifically on valuing that is morally relevant, whereas Doris's account seems to capture, problematically, a broader set of values.

[118] Doris rejects the "lack of values" as an appropriate state for moral evaluation (154-155). For an account like mine, see Sripada: "We sometimes say that an action is expressive not of something a person does care about, but rather what he *fails* to care about, his attitudes of disregard or indifference" (no page numbers)

This indicates that it is important not only to value, but to value the right things or hold the right values as I discussed above.

Finally, and related to the previous point, Doris seems to follow Haidt's suggestion that values and valuing are socially created and motivated, in contrast to morally motivated and created. In the chapter on "Agency" Doris explains how much of our psychological practices—including the expression of values and agency—results in better social acceptance and participation (142-153). In contrast to Doris, I hold that values, or at least the correct values, do not stem from a concern of one's own social well-being, but rather a concern for the well-being of others. Whereas Doris thinks our social nature in large part results in a need to fit in, I hold that our social nature results in concern for others. Andrew Sayer makes the same point:

> As social beings, we simply cannot live without developing some sense of how actions affect well-being and how we ought to treat one another…Almost any kind of social role presupposes shared ideas about how people should properly treat others. The moral values or judgments may not be articulated particularly clearly—often people's dispositions, their practical sense and their 'values-in-use' may be more important than their espoused ideas in guiding what they do—but they can be intelligent and enable people to harmonize their conduct with others and live well…I shall argue that they are related not merely to people's awareness of social norms—of what is deemed proper in their society—but to their understanding of, or feel for, the well-being of those involved. (Sayer 2011, 144).[119]

In short, I am more optimistic than Doris about the genuine—versus socially motivated—moral nature and concerns of individuals (this echoes the way in which I contrast my account to Haidt's).

Conclusion

In this Chapter I have brought together issues raised in Chapter 2 and 3 to offer a view of automatic moral judgment and decision-making that accounts for the empirical—both formal

[119] Sayer also quotes Seyla Benhabib here: "The domain of the moral is so deeply enmeshed with those interactions that constitute our lifeworld that to withdraw from moral judgment is tantamount to ceasing to interact, to talk and act in the human community (Benhabib, 1992, p. 126).

research and our daily experiences—and the normative. By accounting for the empirical, I mean gives adequate consideration to the ample research on implicit processing and attitudes. I have provided an account that fits with many of our intuitions about action and evaluations of moral responsibility—specifically in cases where people appear not to deliberate, but nevertheless seem morally responsible. On the flip side, I have argued—contra Haidt and his interlocutors—that automaticity is not necessarily atavistic or amoral. To say that our moral judgments and decisions are in large part driven by automatic, implicit processes does *not* entail that we are wantons or purely instinctual or even purely socially motivated. I have argued that even when actions are formed solely from our automatic processing, they can still have normative substance. Give this, the empirical research on System 1 processing does not necessarily flatten our moral lives or agency, as suggested by skeptics of Haidt, but instead potentially deepens it.

Works Cited

Anderson, Elisabeth. (1993). *Value in ethics and economics*. Cambridge, Mass.: Harvard University Press.

Arpaly, Nomy. (2003). *Unprincipled virtue: An inquiry into moral agency*. Oxford ; New York: Oxford University Press.

Bargh, J., & Gollwitzer, P. (1994). Environmental control of goal-directed action: Automatic and strategic contingencies between situations and behavior. *Nebraska Symposium on Motivation. Nebraska Symposium on Motivation, 41*, 71-124.

Bargh, J., Gollwitzer, P., Lee-Chai, A., Barndollar, K., Trötschel, R., & Devine, Patricia. (2001). The Automated Will: Nonconscious Activation and Pursuit of Behavioral Goals. *Journal of Personality and Social Psychology, 81*(6), 1014-1027.

Baron, Andrew Scott, and Mahzarin R Banaji. (2006) The Development of Implicit Attitudes. Evidence of Race Evaluations from Ages 6 and 10 and Adulthood. *Psychological Science* 17(1): 53-8.

Bartra, Oscar, Joseph T Mcguire, and Joseph W Kable. 2013. "The Valuation System: A Coordinate-based Meta-analysis of BOLD FMRI Experiments Examining Neural Correlates of Subjective Value." *NeuroImage* 76: 412-27.

Blankenship, Kevin L., Duane T. Wegener, and Judd, Charles M. 2008. "Opening the Mind to Close It: Considering a Message in Light of Important Values Increases Message Processing and Later Resistance to Change." *Journal of Personality and Social Psychology* 94(2): 196-213

De Groot, Judith I.M. I, and Linda Steg. 2010. "Relationships between Value Orientations, Self-determined Motivational Types and Pro-environmental Behavioural Intentions." *Journal of Environmental Psychology* 30(4): 368-78.

Dijksterhuis, Ap, & Aarts, Henk. (2010). Goals, attention, and (un)consciousness. *Annual Review of Psychology, 61*, 467.

Doris, John M. (2015). *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford University Press.

Frankfurt, Harry. (1988). *The importance of what we care about : Philosophical essays*. Cambridge [England]; New York: Cambridge University Press.Jaworska, Agnieszka. (2007). Caring and Full Moral Standing. *Ethics, 117*(3), 460-497.

Frensch, Peter, & Rünger, Dennis. (2003). Implicit Learning. *Current Directions in Psychological Science, 12*(1), 13-18.

Grabenhorst, Fabian, and Edmund T Rolls. 2011. "Value, Pleasure and Choice in the Ventral Prefrontal Cortex." *Trends in Cognitive Sciences* 15(2): 56-67.

Hailey, Sarah, & Olson, Kristina. (2013). A Social Psychologist's Guide to the Development of Racial Attitudes. *Social and Personality Psychology Compass, 7*(7), 457-469.

Held, Virginia. (2006). *The ethics of care: Personal, political, and global*. Oxford; New York: Oxford University Press.

Hitlin, Steven, and Jane Allyn Piliavin. 2004. "Values: Reviving a Dormant Concept." 30: 359-93.

Jaworska, Agnieszka. (2007). Caring and Full Moral Standing. *Ethics, 117*(3), 460-497.

Levy, Dino J, and Paul W Glimcher. "The Root of All Value: A Neural Common Currency for Choice." *Current Opinion in Neurobiology* 22, no. 6 (2012): 1027-38.

Maio, Gregory R. 2010. "Mental representations of social values." *Advances In Experimental Social Psychology,* 42: 1-43.
Maio, Gregory R. and James M. Olson. 1998. "Values as Truisms: Evidence and Implications." *Journal of Personality and Social Psychology* 74(2): 294-311.
Maio, Gregory R., Ali Pakizeh, Wing-Yee Cheung, Kerry J. Rees, and Carver, Charles S. "Changing, Priming, and Acting on Values: Effects via Motivational Relations in a Circular Model. 2009. "*Journal of Personality and Social Psychology* 97(4): 699-715.
Markovits, Julia. (2010). Acting for the Right Reasons. Philosophical Review. 119 (2).
McAdams, Dan P., & Olson, Bradley D. (2010). Personality development: Continuity and change over the life course. *Annual Review of Psychology, 61*, 517
Nichols, Shaun. (2002). How Psychopaths Threaten Moral Rationalism: Is it Irrational to Be Amoral? *The Monist, 85*(2), 285-303.
O'Doherty, John P. 2014. "The Problem with Value." *Neuroscience and Biobehavioral Reviews* 43: 259-68.
Rangel, Antonio, and John A Clithero. 2012. "Value Normalization in Decision Making: Theory and Evidence." *Current Opinion in Neurobiology* 22(6): 970-81.
Ravlin, E. C., & Meglino, B. M. 1987a. Issues in work values measurement. In L. Preston (Ed.). Research in corporate social performance and policy (Vol. 9, pp. 153–183). JAI Press Inc.
Ravlin, E. C., & Meglino, B. M. 1987b. Effect of values on perception and decision making: A study of alternative work values measures. Journal of Applied Psychology, 72, 666–673.
Snow, Nancy. 2010. *Virtue as social intelligence: An empirically grounded theory*. New York: Routledge.
Sayer, Andrew. 2011. *Why things matter to people: Social science, values and ethical life*. Cambridge, UK; New York: Cambridge University Press.
Shoemaker, David. 2003. Caring, Identification, and Agency. *Ethics, 114*(1), 88-118.
Smith, Angela M. 2005. Responsibility for Attitudes: Activity and Passivity in Mental Life. *Ethics*. 115 (2).
Sripada, Chandra. 2015a. Self-expression: a deep self theory of moral responsibility. *Philosophical Studies* Published online August 12, 2015 doi 10.1007/s11098-015-0527-9
Sripada, Chandra. 2015b. Acting from the gut: Responsibility without awareness, *Journal of Consciousness Studies*, 22 (7–8).
Sripada, Chandra. 2015c. Moral Responsibility, Reasons, and the Self. In *Oxford Studies in Agency and Responsibility* (p. Oxford Studies in Agency and Responsibility, Chapter 12). Oxford University Press.
Svavarsdóttir, Sigrún. 2014. Having value and being worth valuing. *The Journal of Philosophy* 111(2): 84-109.
Twain, Mark. 1992. *The Adventures of Tom Sawyer & The Adventures of Huckleberry Finn*. Ware: Wordsworth Classics, 1992.
Torelli, Carlos J., Andrew M. Kaikati, and Carver, Charles S. 2009. "Values as Predictors of Judgments and Behaviors: The Role of Abstract and Concrete Mindsets." *Journal of Personality and Social Psychology* 96(1): 231-47.
Wilson, Timothy. (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge, Mass.: Belknap Press of Harvard University Press.
crucial (Murdoch 1970, 36; cf Sayer 2011, 97).

CHAPTER 5

At the beginning of the dissertation, I introduced two characters: Steve and Debbie. Recall that when Steve and Debbie were asked about a moral question, they both made the same moral judgment—the war in Afghanistan was morally wrong—but when asked for reasons, gave different responses. Steve articulated something of a utilitarian view, while Debbie insisted: "It's just wrong." Up to this point, I have attempted to establish, first, that it is empirically plausible that Debbie's moral judgment genuinely did not stem from conscious deliberation and, second, that her moral judgment can hold just as much moral worth as Steve's deliberative judgment. In other words, if Steve is morally responsible for his good moral judgment, Debbie is morally responsible, too. As I explained in Chapter 4, this is because in both Steve and Debbie's cases, moral reasons are being invoked via moral values. The difference is that Steve's moral reasons and values are acquired, triggered, and executed via System 2, while Debbie's are acquired, triggered, and executed via System 1.[120]

In this final chapter, I discuss two major worries one might have about my account of value-guided automaticity:

1. In some cases, actions will be driven by conflicting moral values. How ought these situations be morally evaluated on my account? Relatedly, sometimes, having the right value will lead to a morally bad moral judgment or decision. How can my value-based account handle this?
2. While it may seem fine to *praise* those who act without control or awareness,[121] it is not as clear that it is appropriate to *blame* those who act without control or awareness.

[120] It isn't actually clear that Steve's moral reasons or values are *acquired* via conscious deliberation, and it is worth exploring what reflectivists would thus say about the normative status of his moral judgment if his moral reasons are triggered and executed consciously, but not acquired via conscious deliberation.

[121] I say act "without control" because the agents I discuss in chapter 4 did not make a conscious decision at any point that has led to the relevant action. I am not focused on cases, for example, where one consciously chooses to be more patient and over time nonconsciously acts more patient. Instead, I am interested in cases where the agent does not choose to acquire particular values, nor act upon them. Thus, there is a sense in which it seems that the agents act on particular values without not only awareness, but also control. This is precisely why many think that individuals are not responsible for having or acting upon implicitly acquired and held stereotypes. It is often said "but I didn't *choose* to have those implicit biases, so how can I be responsible for them?"

Does my account thus have unjust or unfair implications? Relatedly, does my account rely too heavily on moral luck?

I now turn to each of these issues, aiming to further clarify my account. At the end of the chapter, I discuss areas for future research.

Objection #1: Conflicting values and good values gone bad

In some cases, actions will be driven by conflicting moral values. How ought these situations be morally evaluated on my account? Relatedly, sometimes, having the right value will lead to a morally bad moral judgment or decision. How can my value-based account handle this?

In my analysis in previous chapters, I have focused on somewhat simple cases, where I have described someone as acquiring *a* value that is automatically triggered and executed. However, the process will rarely be so simple. We hold *many* values, some that are consistent with each other and some that conflict. I care about my health and spending time outdoors; those complement each other well. On the other hand, I might simultaneously care about excelling at my job and spending more time with family and these values may sometimes be in tension with each other. In reality, we hold a complex web of values, cares, and concerns. This has several implications for my account.

First, there could be multiple values that guide any particular moral judgment and decision. For example, one might decide to volunteer at the animal shelter because she values giving back to her community *and* helping animals. Both values play a role in the moral decision. This in and of itself does not seem to pose a problem for my account as both values here—giving back and helping animals—entail good/right reasons for acting. It might be said, however, that one who acts for both reasons is more morally praiseworthy than the person who only acts for one. If for example, Jordan volunteered because he really wanted to help animals, but didn't care about giving back to his community, we might say he is *less* praiseworthy than Selah who cares about both, though nevertheless praiseworthy. Nomy Arpaly introduces the idea

of "degrees" of praiseworthiness and blameworthiness in her account of unprincipled virtue. She argues, for example, that A's action is more morally praiseworthy than B's action if A's moral concern is deeper:

> *Praiseworthiness as Responsiveness to Moral Reasons (revised version):* For an agent to be morally praiseworthy for doing the right thing is for her to have done the right thing for the relevant moral reasons—that is, for the reasons for which the action is right (the *right reasons* clause); and an agent is more praiseworthy, other things being equal, the deeper the moral concern that has led to her action (the *concern* clause). Moral concern is to be understood as concern for what is in fact morally relevant and not a concern for what the agent takes to be morally relevant. (2003, 84)

In the language of my account, an agent is more praiseworthy, other things being equal, the deeper the care or valuing is that led to her action. The parent who shows up to the PTA meeting because they care deeply would be more praiseworthy than the parent who shows up because they care somewhat.[122] Both parents, however, would be praiseworthy on my account because their action was motivated by the right moral values, cares, or reasons.[123]

More complicated cases are those in which our actions are guided by incongruent or conflicting values. Imagine that the parent attends the PTA meeting both because she cares about her child *and* because she wants to make her ex look bad by being more involved than her. Caring about one's child is good moral reason for attending the meeting, whereas making the ex look bad is not. The question arises whether the parent in this case is morally praiseworthy for

[122] Remember that, on my account, the parents need not be consciously aware of the existence or depth of their caring. They might think they don't care much, or think they care deeply, yet find that their actions or responses suggest otherwise.

[123] Arpaly argues that depth of concern is often determined by whether one would have acted under more difficult circumstances. This conversation about counterfactual cases—"Would she have helped out even if she was depressed?"—is unnecessary for Arpaly's point and raises concern. Markovits (2010) discusses the issue, explaining that the evaluation of one's action should not be influenced by whether one would have done the same under different circumstances. We might evaluate someone's *character* differently depending on how consistent they are in their actions or commitments, but it does not seem appropriate to evaluate *individual actions* upon counterfactuals (210-213). However, while I find Markovits's point about the consideration of counterfactuals compelling, it does not undercut Arpaly's point about the significance of *depth* of concern, on which I think Arpaly is right. Dana Nelkin (2014) also discusses Arpaly's "depth of concern" and writes that difficulty in performing an action may sometimes, but not always indicate a depth of concern (6-7)—further suggesting that Arpaly's analysis needs revising/further clarification. I discuss this latter point in more detail in the next section.

her attendance. The best answer, I think, is that the parent is *both* praiseworthy and blameworthy. She is praiseworthy for acting out of concern for her children *and* blameworthy for acting out the desire to bring her ex down. Note that this doesn't necessarily mean we would both praise and blame her. Remember that whether someone is praise/blameworthy and whether we should, in practice, praise/blame them are distinct issues. Thus, it does not seem incoherent to evaluate someone as both praise- and blameworthy.[124] Given that some actions have multiple motivating reasons and what matters for moral evaluation is whether the action's motivating reasons are good or bad, we should expect to have complex moral evaluations. Note also that this question is not unique to my account of value-guided automaticity. Many reasons-based accounts in which deliberation is necessary for moral responsibility will also have to explain how to evaluate actions guided by multiple or conflicting reasons. In other words, my automaticity-focused account does not introduce a new problem.

Consider a different kind of case, where one acts for good moral reasons, or, on my account, the right moral values, but does the wrong thing. Imagine that Rwan's young adult child has committed a crime that has harmed others, and if convicted will be sent to prison. Because Rwan cares deeply about the child, Rwan does whatever possible prevent her from being convicted, deciding even to lie under oath.[125] Presuming that lying under oath is the morally wrong thing to do, how ought we evaluate Rwan's action on my account? As with the case in the preceding paragraph, the appropriate evaluation in Rwan's case might involve both praise and

[124] Arpaly also discusses cases in which an agent has "mixed motives" (2003, 112-114). She gives the example of a young man who joins the Nazi party in large part out of concerns of social justice, but also in part out of frustration and hatred. She describes his motives as a "morally mixed bag." While Arpaly acknowledges that these kinds of complex cases exist, she does not give a clear answer about how such agents or actions should be morally evaluated. She concludes instead: "…as moral worth depends on the complex matter of people's motives, diagnosing moral worth in real life can be very difficult…" (114). We might guess, however, given her standard of praiseworthiness as responsiveness to moral reasons, that she would either agree with my claim that such persons are both morally praiseworthy and blameworthy or argue that people with complex motives are praiseworthy or blameworthy to a lesser degree. See also Sripada 2015.

[125] Thanks to Aaron Spink and Bill Talbott for pushing me on this kind of case.

blame. We might blame Rwan for failing to properly value the justice system and the suffering of those who were harmed, while simultaneously praising Rwan for showing such deep concern for her child. While it might seem counterintuitive to praise Rwan in this case, we could easily imagine some of Rwan's friends and neighbors claiming that Rwan did an admirable thing if not the right thing. Thus, it seems plausible that Rwan's action warrants both praise and blame.

This suggestion, however, conflicts with my earlier cases and definition of praiseworthiness. Recall the "right reasons clause" of Arpaly's praiseworthiness thesis: "For an agent to be morally praiseworthy for doing the right thing is for her to have done the right thing for the relevant moral reasons—that is, for the reasons for which the action is right." In the case of lying under oath, Rwan has not done the right thing for the right reason, but instead done the wrong thing. Two questions thus arise: can one do the wrong thing for the right reason? And if so, can one be praiseworthy for doing the wrong thing?

As I have mentioned above, an action can be guided by multiple moral reasons or moral values. Thus, it seems possible that one can do the wrong thing for the right reason, though not *purely* a right reason. Doing the wrong thing will also involve a bad moral value or a lack of relevant moral values.[126] This is again how we can describe the case of Rwan. Rwan's lying is motivated both by care for her child, but also by a lack of appropriate care for justice or the suffering of the wronged.[127] Arpaly (2003) describes a scenario where a young, uneducated, and

[126] Excluding cases where one does the wrong thing because of confusion or epistemic lacking. If I, new in town, did not realize there were *two* karaoke bars, and thus despite my best planning show up very late to your birthday party, I may have done the wrong thing without having any bad motivations or lack of relevant motivations.

[127] It may seem odd that I describe Rwan's lack of "care for justice" as a failing given that I above rejected the idea, following Markovits, that one must act out of "concern for the right, or duty," for example. I do not mean to suggest here that one *must* act out of concern for the right, but rather that in this particular case, it would seem reasonable to expect that Rwan care about fairness or justice. In lots of the other cases I discuss—caring for my sister, Autrey's jumping on the subway tracks, the Holocaust rescuers—I still reject the idea that these actors need act out of "concern for the right" or "concern for duty." Instead I have argued that acting simply because they care about others, whether they see that as morally right or required, is sufficient for moral agency and in turn moral praise. Thus, I do not think concern for justice or concern for duty is necessary in the sense that Kant and others might—I

naïve German man joins in the Nazi army while it is still a fringe group. Arpaly explains that this young man (Hans Miklas) is in large part motivated by concerns of social justice and national honor. However, it is only because these concerns are conjoined with frustration and hatred that Hans is susceptible, or drawn to, the Nazi's anti-Semitic propaganda (Arpaly 2003, 112-113). In the cases of Rwan and Hans, it seems problematic to suggest that they are wholly blameworthy given their morally admirable values. This is perhaps illustrated by the fact that it seems wrong to say that Rwan and Hans are as blameworthy as Rwan*, who lies under oath simply because she doesn't like punishment and is callous about victim's suffering,[128] or Hans*, who joins the Nazi army simply because he hates Jews. We are inclined to give Rwan and Hans less harsh judgment, or even feel sympathetic, despite our certainty that they have acted wrongly and in some sense deserve blame. My account of value-guided automaticity actually allows for a nuanced analysis of Rwan and Hans's actions because it highlights that agents can have complex values which explains our sometimes complicated evaluations of people's actions.

The same analysis can be applied to our actors who do morally praiseworthy things that might invoke complicated intuitions. For example, I have described Wesley Autrey's action as heroic and motivated by a deep concern for others' well-being. However, it is also worth noting that Autrey jumped on the track in front of his two daughters, and in putting himself at such risk may have caused his daughters to witness his death. For this, it seems that Autrey may not have shown the proper concern of his daughters' well-being. Again, my value-guided account allows us to give a more complex evaluation of Autrey's action given this fact. We might still say that

---

think acting from inclination in many cases will warrant praise and lack of inclination in some cases will warrant blame. Again, following Markovits, I hold only that one must act for a noninstrumental moral reason. In some cases, the relevant moral reason will involve justice, fairness, and duty (e.g. "because I made a promise"), but in many cases—I have tried to argue—the relevant moral can be "because I wanted to help," "because I felt moved," "because I care about her," etc.

[128] Think here about the recent phenomenon of "affluenza parents": http://www.democracynow.org/2014/2/7/affluenza_defense_lands_wealthy_teen_in

Autrey is admirable for his heroic action, while simultaneously expressing concern about the situation in which he put his daughters. His complex values or depth of concern results in a complex evaluation.[129] The same may be said of the individuals who helped Jewish refugees. While we again think they expressed morally admirable action, it would not be amiss to point out that in many cases, rescuers put their family's and neighbor's lives at risk. It would not seem misguided if a wife begged her husband to refuse a Jewish family stay and blamed him when the SS did in fact find the refugees, which resulted in harm to their children. She might even resent him for putting the family at risk even in the case that the refugees are never discovered.[130] In sum, not all actions will either be straightforwardly praiseworthy or blameworthy. This does not necessarily indicate a problem with my account, however; in fact, it may explain conflicting intuitions we have about a variety of moral actions.

Objection #2: Praise/blame asymmetry

While it may seem fine to praise those who act without control or awareness, it is not as clear that it is appropriate to blame those who act without control or awareness. Does my account thus have unjust or unfair implications? Relatedly, does my account rely too heavily on moral luck?

Many of the cases I present in my analysis involve actors or actions that ought to be praised. I focus in large part of cases of praise because one of the main aims of my project is to show that automatic processing deserves more credit than it currently receives in the philosophical literature. Thus, I attempt to show that cases in which people do genuinely admirable things

[129] Note that I maintain this is all happening at the subconscious level—outside of his conscious awareness.
[130] I am less sure about what to say about the person who does not aid the Jewish refugees because of the fear that she herself will be killed. But ultimately, whether this action is praise- or blameworthy or both will depend on ethical theory. On some ethical theories, fear of self-harm may justify refraining from acting to help another, on other ethical theories, it will not. Given my theoretical leanings, I believe care ethics may be particularly helpful in illuminating the moral obligation in such cases where our needs for care and protection come into conflict with others' needs for care and protection. But I do not attempt to resolve the moral evaluation here, but rather recognize it as a more difficult, gray case, one that involves conflicting moral values (care for others vs. care for oneself).

without deliberating—jump on subway tracks, illegally take in refugees, lie to help a friend—can be morally praiseworthy because moral values and reasons are deployed through the automatic system. In these cases, I argue, actors may not have deliberated *at any point*, but are nevertheless morally praiseworthy. On the flip side of such analysis, however, is the implication that we ought to also evaluate actors and actions and blameworthy in cases when people do bad things without deliberating *at any point*. This idea, however, is quite counterintuitive. We often think that if someone did something wrong without knowing, consciously intending, or choosing it, they are not blameworthy. Is it possible, then, that my account only applies to praiseworthy, and not blameworthy, actions? And if so, how can I account for the asymmetry? In this section, I will draw upon a well-discussed kind of case relevant to this objection crafted by Susan Wolf (1987) and try to pinpoint the precise force of this objection by exploring whether it is a lack of capacity or ability that seems problematic. Ultimately, I will show that while Wolfian analysis of these kinds of cases are intuitively plausible, it is ultimately difficult to show that people who have not chosen their values are necessarily shielded from evaluations of blame, and thus the asymmetry does not actually exist.

To cash out intuitions about blameworthiness more, consider a case often discussed in the literature on moral responsibility. JoJo is the son of an evil dictator who himself grows up to be an evil dictator.[131] As a child, JoJo is a favorite of his father's and is thus given special attention and spends a significant amount of time with his father. JoJo grows up somewhat in the image of his father, sharing many of his same values. This results in JoJo, like his father, sending people to prison and torturing at whim. JoJo is not coerced into this behavior; he genuinely desires to act

[131] This case is introduced by Susan Wolf (1987) and widely discussed in the literature on moral responsibility and blame. See, for example: Applebaum 1997, Greenspan 2003, Law 2003, Levy 2003, Vincent 2011, Talbert 2012, Maibom 2013, Bedrick 2014, Ciurria 2014, Faraci and Shoemaker 2014, Mason 2015, and Shoemaker 2015.

this way and doing so accords with his values. Wolf describes what she takes to be a common intuition about evaluating JoJo:

> In light of JoJo's heritage and upbringing—both of which he was powerless to control—it is dubious at best that he should be regarded as responsible for what he does. It is unclear whether anyone with a childhood such as his could have developed into anything but the twisted and perverse sort of person that he has become. (1987, 54)

People also raise this worry in a modern American context. Imagine that Bob was raised in a small town in the South, where all throughout his childhood and into his adulthood his friends and family espoused explicitly racist views. These racist views were reinforced by the media that Bob was exposed to, teachers and religious leaders in the community, and the various social institutions in the community. As a result, Bob has absorbed implicitly held racist values and thus behaves in racist ways—sometimes deliberately but sometimes due to nonconscious processing. Given that it seems Bob has not himself chosen these racist attitudes and really could not have come to hold other values, many have the intuition that Bob is not blameworthy.[132] Wolf writes, for example:

> "…this new proposal explains why we give less than full responsibility to persons who, though act badly, act in ways that are strongly encouraged by their societies—the slaveowners of the 1850s, the Nazis of the 1930s, and many male chauvinists of our fathers' generation, for example. These are people, we imagine, who falsely believe that the ways in which they are acting are morally acceptable, and so, we may assume, their behavior is expressive of or at least in accordance with these agents' deep selves. But their false beliefs in the moral permissibility of these actions and the false values from which these beliefs derived may have been inevitable, given the social circumstances in which they developed. If we think that the agents could not help but be mistaken about their values, we do not blame them for the actions those actions those values inspired?" (56-57)[133]

[132] Thanks to Sara Goering and Aaron Spink for pushing me on this kind of case.

[133] Wolf calls this a hypothetical—perhaps they *are* able to recognize that their actions are wrong, in which case they *are* morally responsible. So, her suggestion here is that *if* they are not able to recognize that their actions are wrong, *then* they are not morally responsible. She thinks this is an empirical question, one that is hard to answer. I will answer it.

These characters—the slaveowners, the Nazis, and male chauvinists—could not help but have morally problematic views, like Bob and JoJo.[134]

But what really pumps the intuition that folks like JoJo, Bob, and slaveowners are not morally blameworthy for their attitudes and actions? It cannot simply be that they did not choose their attitudes, values, and beliefs, because most of us did not choose our attitudes and beliefs. We are just as much the product of our environment as JoJo and Bob are. Wolf articulates the point:

> Our judgment that JoJo is not a responsible agent is one that we can make only from the outside—from reflecting on the fact, it seems, that his deepest self is not up to him. Looked at from the outside, however, our situation seems no different from his... it is not up to any of us to have the deepest selves we do. (54)
>
> If JoJo is not responsible because his deepest self is not up to him, then we are not responsible either. (54)

Very likely, you have come to hold your egalitarian beliefs and commitments because friends, family, or teachers taught you that equality and compassion are important values. In turn, you have grown into a person who cares about equality and the well-being of others. In fact, this *must* be the narrative about how you have acquired your values. For, if it were the case that you were actually taught oppressive and unjust values, but somehow came to recognize the importance of equality and compassion, then it *would* seem possible for JoJo, Bob, the slaveholders, and Nazis to have cultivated egalitarian values despite their upbringing.[135] So, the corollary of the idea that JoJo et al. could not have chosen other values or developed other deep selves is the idea that those with egalitarian commitments also could not have chosen other values or developed other

[134] Michele Moody-Adams (1994) calls this the "inability thesis": "…that sometimes one's upbringing in a culture simply renders one unable to know that certain actins are wrong," citing Susan Wolf, Alan Donagan, and Michael Slote as proponents of the thesis (292-293).

[135] One might be inclined to clarify that while not impossible for JoJo et al. to hold egalitarian values, it is *more difficult* because of their social and moral education. I will say more about how difficulty affects evaluations of moral responsibility below.

deep selves. And thus, if choosing our values is necessary for moral responsibility, none of us is morally responsible.[136] Recall from my discussion in Chapter 4 about the claim that our values can be acquired implicitly, Shoemaker's claim that none of our current cares are under our voluntary control at the current time (Shoemaker 2003, 105). We cannot start or stop caring about something at will and when we do change our cares it is often because other cares have evolved or take priority. Given this, the claim that JoJo is not morally responsible because he does not choose his moral values (or lack thereof) while we are morally responsible because we choose ours seems implausible.[137]

It might be objected here that the person with the egalitarian commitments *could* have chosen otherwise because she has the *capacity* to reflect on reasons and moral commitments. The importance of capacity is discussed in depth in the moral psychology literature on psychopaths.[138] A common intuition about psychopaths is that they are not morally responsible for their actions not because they didn't choose their values, but because they appear to lack the *most foundational capacities* that enable one to identify right from wrong.[139] It is hypothesized

[136] If one insists here that we *do* have the ability to reflect upon and choose our values, whereas JoJo does not, note that even that ability or inclination to reflect and choose will be determined by our environmental circumstances, as Wolf explains: "Though I can step back from the values my parents and teachers have given me and ask whether these are the values I really want, the "I" that steps back will itself be a product of the parents and teachers I am questioning" (Wolf 1987, 52). Thus, if being influenced by external forces undermines JoJo's agency, ours is also undermined given that we also cannot escape external influences.

[137] In earlier work (1980) Susan Wolf argues that there is an important asymmetry here. She argues that one *cannot be blameworthy* if her bad actions are determined, but one *can be praiseworthy* if his right actions are determined: "Determination, then, is compatible with an agent's responsibility for good action, but incompatible with an agent's responsibility for a bad action. The metaphysical conditions required for an agent's responsibility will vary according to the value of the action he performs" (158). I will not investigate Wolf's analysis further here because I agree that even when the good-agent's values are determined, he is morally responsible. Wolf and I thus disagree on whether the bad-agent is morally responsible when her values are determined, but I will argue below that the cases in which Wolf thinks an agent's values are determined are not actually so. It is also worth noting that Wolf and I have different audiences: she is trying to convince determinists that we can be responsible for determined action that is morally right, while I am trying to convince indeterminists that we can be responsible for morally wrong action that seems determined.

[138] See, for example: Greenspan 2003, Levy 2005, Talbert 2008, Shoemaker 2011, Vincent 2011, Watson 2011, Scanlon 2012, Watson 2013, Bedrick 2014, and Nelkin 2015.

[139] See Maibom 2008, for a dissenting view. I do not take a stance here on whether psychopaths are in fact morally responsible. My goal is simply to point out that if someone holds that psychopaths are not morally responsible, it is

that because psychopaths lack the capacity for empathy they are not able to recognize moral reasons or the force of moral reasons. They thus, conflate moral rules and conventional rules—where a moral rule is something like *don't hit others* and a conventional rule is something like *don't eat while standing*. Psychopaths see both moral rules and convention rules as social norms, rather than moral norms. Thus, for psychopaths moral rules are to be respected merely because society tells us to do so—there is no force to moral rules beyond this. Children as young as the age of 6 think that conventional rules are okay to break if given permission, whereas moral rules are *not* okay to break, even when given permission. While typical-functioning children recognize the distinction at a rather young age, psychopaths don't understand this distinction because of, it is suggested, a failure of their Violence Inhibitor Mechanism (Blair et al. 1995). Given this lack of the capacity to respond to moral reasons, many hold that psychopaths are not morally responsible. The same may be said about less-intelligent animals, robots, and infants.

However, it does not seem accurate to describe JoJo, Bob, and others as lacking the *capacity* to respond to moral reasons. Their moral failing is not the same as the psychopath's. It is presumed that their cognitive mechanisms are fully functioning, but they have been given misinformation or bad models to follow. So perhaps it is more accurate to say that JoJo et al. do not have the *ability* to understand moral reasons and moral commitments. "Ability" is distinguished from "capacity" insofar as one's ability may be affected by external, environmental factors, whereas we would expect one's capacity to stay consistent across contexts or situations—I might have the capacity to play tennis, but not the ability given that I left my rackets at home. Wolf (1987) seems to endorse this latter view: though she argues that JoJo et al. are insane, this seems to be a software rather than hardware failing. By this, I mean that JoJo et

---

likely because they lack the *capacity* to distinguish between right and wrong. I then explore whether this is an appropriate description of JoJo et al.

al. have not been instilled with the proper values and thus cannot distinguish between right and wrong, even though their brain structure may be the same as yours and mine. Wolf explains that the problem in JoJo's case it that he was given a highly problematic moral education and thus has problematic desires and values. But these desires and values function in "normal" integrated, expressive way—JoJo's actions are structured in the same way ours are (unlike the psychopath or hypnotized person) (368). Because JoJo was not taught otherwise, he lacks the ability to know right from wrong. Wolf further claims that people lacking the ability to know right from wrong are not morally responsible: "Since these characters lack the ability to know right from wrong, they are unable to revise their characters on the basis of right and wrong, and so their deep selves lack the resources and the reasons that might have served as a basis for self-correction" (58).[140] People who have the ability to know right from wrong, in contrast, can revise their characters—for better or worse—and thus are morally responsible.

I think Wolf has pumped a very plausible intuition here—that we can't hold people morally responsible if they don't have the ability to know right or wrong. However, it is not obvious that the characters above do in fact lack the ability to know right from wrong, in general.

[140] Wolf also makes the point in her essay "Asymmetrical Freedom" (1980): "Yet it seems he ought not to be blamed for committing his crime, for, from his point of view, one cannot reasonably expect him to see anything wrong with his action. We may suppose that in his childhood he was given no love—he was beaten by his father, neglected by his mother. And that the people to whom he was exposed when he was growing up gave him examples only of evil and selfishness. From this point of view, it is natural to conclude that respecting other people's property would be foolish. For presumably no one had every respected his" (159-160). Note, however, that this character is significantly different than JoJo et al. In the backstories of JoJo et al., the relevant fact is that the characters have been taught the wrong beliefs and values. In this case that Wolf describes, the character has been beaten and neglected. It is not simply a matter of bad moral education. Instead, this beaten and neglected character likely has cognitive and psychological damage. Psychologist Darcia Narváez explains, for example: "When early care is poor, it sets us on a suboptimal trajectory and leads to a different becoming. Undercare of our evolved needs in early life leads to deficiencies in the brain of structural integrity, hormonal regulation, and system integration that lead to sociality. Early stress is especially toxic to long-term well-being because it undermines development of brain and body systems. When a child does not receive appropriate care, the more primitive brain systems may dominate social relations, curtailing optimal moral growth. Stress reactivity will overwhelm psychological and moral functioning" (2014, 126). Given this, the character Wolf describes here in "Asymmetrical Freedom" is more akin to the psychopath than JoJo. The failing here is not simply lack of moral education, but neglect, which can have irreversible cognitive and psychological effects.

Wolf extracts the idea that they lack the ability to know right and wrong from their heinous actions: "Sanity, remember, involves the ability to know the difference between right and wrong, and a person who, even on reflection, cannot see that having someone tortured because he failed to salute you is wrong plainly lacks the requisite ability" (1987, 56). But there are two points to make about her inference: first, lots of people who commit heinous acts do actually have the ability to know right from wrong; they likely are able to distinguish right from wrong in many other contexts. Take JoJo's evil dictator father, Jo, for example. While Jo treats many people abhorrently, he is described as showing particular affection and care for JoJo, which is morally admirable (though he certainly fails in providing JoJo an appropriate moral education). He might think it wrong for people to harm his child or that he ought to show particular care to JoJo and raise him to be a strong and competent leader. It is likely the case that the other characters Wolf describes also care about moral concepts and values in other contexts; for example, they might care about things such as loyalty/friendship, compassion (for example, killing a suffering animal to put it out of its misery), and fairness. Think of Walter White from AMC's *Breaking Bad*. Walt commits heinous acts throughout the show, yet always maintains a strong sense of commitment to his family—in fact, he claims to commit these egregious acts so that he can provide a better life for them. JoJo et al. are probably similarly morally complex; morally confused in particular contexts—involving dissenting villagers, Black people, slaves, Jewish people, and women—but not amoral all together. Hence, we cannot infer from their egregious actions that that they are unable to understand right and wrong full stop; this seems unlikely unless we are talking about genuine psychopaths, which Wolf is not doing. She is not talking about people who lack cognitive capacities, but rather cannot see right and wrong due to their cultural upbringing.[141]

[141] Faraci and Shoemaker (2010) make a similar point: "What JoJo lacks is the knowledge of what *is* right and what *is* wrong, not the difference between them. In addition, it remains unclear that JoJo's lack of knowledge stems from

Furthermore, if Wolf's claim that "…a person who, even on reflection, cannot see that having someone tortured because he failed to salute you is wrong plainly lacks the requisite ability" is true, most actors committing heinous crimes lack the requisite ability and are thus not blameworthy. For example, one might also say that "a person who, even on reflection, cannot see that torturing and killing another simply because he is attracted to persons of the same sex is wrong plainly lacks the requisite ability," or "a person who, even on reflection, cannot see that see that beating his wife simply because she accidentally broke a plate is wrong plainly lacks the requisite ability," or "a person who, upon reflection, cannot see that having sex with someone who is unconscious is wrong plainly lacks the requisite ability," etc. In other words, Wolf implies that any time someone commits a clearly abhorrent action, it is likely (or in her words: "plainly") because they do not understand right from wrong. If true, our intuitions about and practices of blame would have to change radically; sex offenders, wife-beaters, and gay-bashers may in most cases lack moral responsibility. This appears problematic as these are perhaps paradigmatic cases of blameworthiness. Wolf acknowledges this possible implication of her account and comments optimistically that any egregious act or mistaken moral belief may not necessarily be indicative of the inability to distinguish right from wrong (61). She writes, for example

> …when someone acts in a way that is not in accordance with acceptable standards of rationality and reasonableness, it is always appropriate to look for an explanation of why he or she acted that way. The hypothesis that the person was unable to understand and appreciate that an action fell outside acceptable bounds will always be a possible explanation… Typically, however, other explanations will be possible, too—for example, that the agent was too lazy to consider whether his or her action was acceptable, or too greedy to care… Other facts about the agent's history will help us decide among these hypotheses. (61).[142]

---

a disability (the disability essential to insanity): just because he does not know that expressions of ill will are immoral, that does not mean that he lacks the ability to know it" (328).

[142] I'm not quite sure how to reconcile her point here with the statement above that JoJo "plainly lacks the requisite ability."

Here, Wolf seems to back away from the claim that JoJo "plainly lacks the requisite ability" to distinguish right from wrong as indicated by his actions. The question that arises then is what facts about the agent's history we would need to know to determine whether the agent genuinely lacks the ability to distinguish between right and wrong. I suggested in the preceding paragraph that when we do look at JoJo et al.'s history, we are likely to see stories in which an individual acts egregiously in some contexts, but not all. Jo the senior is a loving father, the Nazi shows deep care for his community or country, racist Bob wants to help his nephew finish high school and find a good job, the slaveholder thinks that it is important to maintain integrity in his business dealings.[143] Given these likely complex histories and stories, it does not seem accurate to say that these characters lack the ability to know right from wrong. Instead, it should be said that they lack the ability to know right from wrong in a *particular context*.[144] The question then becomes whether these characters are able to understand that they are wrong about the particular relevant issue, not whether they lack the ability to distinguish from right and wrong in general.

To determine whether one is able to know that their action is wrong, we need to know if there is an inferential route that one could follow from their beliefs and values to the recognition

[143] See, for example, Hannah Arendt's (1963) description of Adolf Eichmann: "Half a dozen psychiatrists had certified Eichmann as 'normal.' 'More normal, at any rate, than I am after having examined him,' one of them was said to have exclaimed, while another had found that Eichmann's whole psychological outlook, including his relationship with his wife and children, his mother and father, his brothers and sisters and friends, was 'not only normal but most desirable.' And, finally, a minister who paid regular visits to him in prison after the Supreme Court had finished hearing his appeal reassured everybody by declaring that Eichmann was 'a man with very positive ideas.'" Eichmann, in fact, seem to be something of a Deontologist—holding that one should not make exceptions for himself and should not let emotions or personal connections interfere with moral duties: "An 'idealist' was a man who *lived* for his idea (hence he could not be a businessman, for example) and who was prepared to sacrifice for his idea everything and, especially, everybody. When he asserted during the police examination that he would have sent his own father to his death if that had been required, he did not mean merely to stress the extent to which he was under orders, and ready to obey them; he also meant to show what an 'idealist' he had always been. Of course, the perfect 'idealist,' like everybody else, had his personal feelings, but if they came into conflict with his 'ideal,' he would never permit them to interfere with his actions."

[144] This qualification may make Wolf's comment that JoJo "plainly lacks the requisite ability" more plausible. While we might not have good reason to think that JoJo lacks the ability to distinguish between right and wrong in general, his actions may indicate that he lacks the ability to distinguish between right and wrong when it comes to treatment of his subjects. I will argue, though, that even once qualified in this way, Wolf's inference is not warranted.

that the action under question is wrong.[145] In the cases of JoJo et al., it does seem that there is an inferential route from their values and beliefs to the recognition that their actions are wrong. Jojo et al. understand that their actions cause pain and suffering to others. Presuming they are not psychopaths and have normal cognitive and physiological responses to pain in themselves and others, they are able to understand that they inflict pain.[146] Given that they understand pain and suffering as something to be avoided, they should be able to see that there is harm done in inflicting pain and suffering on their victims. This opens up the inferential route from their beliefs and values to recognition that their actions are wrong. We think that beings who can experience pain and suffering and can recognize it in others should be inclined to treat them well.[147] This is why it seems particularly egregious when persons from marginalized groups fail to recognize the plight of other marginalized groups. When a white woman (Susan) espouses racist views or a Black man (Robert) expresses homophobia, we are particularly shocked[148] because we expect them to be able to easily infer that people in other marginalized groups experience great and pain and suffering, just as Susan and Robert have experienced suffering because of their membership in marginalized groups. Even if Susan and Robert have been taught racists and homophobic views, we are surprised that they are not able to change those views given the straightforward inferential route from their experiences to others'. Likewise, it seems that there is an inferential route from the fact that JoJo et al. experience pain and suffering and

[145] See Heidi Maibom (2013) for what I take to be a compelling argument for such a point. Note also that I do no think this inferential route necessarily be conscious or explicit. I suspect that in many cases one implicitly infers from situation A that situation B is also morally problematic. Maibom does not distinguish between implicit and explicit inference, though I would guess she has explicit inference in mind.

[146] Evidence from brain scans shows that psychopaths do not seem able to recognize pain in others, presumably due to their lack of the ability to empathize. Barring psychopathy or trauma-induced mental disorders, people are able to recognize distress in others.

[147] In fact, I think our intuitions change if JoJo is instead Josephine. I suspect we would expect Josephine to see that her actions are wrong, despite the strong teachings from her father Jo. This could be because we think the inferential route is easier to take for Josephine, given that she likely experiences sexist discrimination, or because of our gendered expectations that women be more empathetic. A problem arises at any rate that we expect Josephine to recognize the wrongness of her actions but not JoJo.

[148] And may think Susan and Robert are *more* blameworthy, though I leave that question aside for now.

presumably want those they care about to be spared it to the fact that their victims also experience pain and suffering and want to avoid it.[149]

However, knowing that I cause others pain and suffering itself does not imply that my action is wrong. JoJo et al. most likely view their harming others as justified—for reasons such as that the villager challenged JoJo's authority, Black people are intellectually and emotionally inferior, Jewish people are ruining the country, or women will create more problems if not put into their proper place. These characters have false beliefs that make them believe they are not acting immorally.[150] In other words, they do not properly recognize right and wrong in the particular context. Are they *able* to recognize right and wrong in the particular context? I do not see what makes them *unable*. Simply that they were taught and modeled the opposite? But people are taught false/incorrect facts and morals often and seem able to recognize error—or acquire better values—despite being taught otherwise. This is precisely what happens in the case of Huck Finn. Huck is indoctrinated with morally problematic values, but in spending time with Jim, acquires better, morally praiseworthy values. He is able to infer (implicitly) new implicitly held beliefs and values from his existing beliefs and values about friendship and affection.[151] Likewise, it does not seem impossible that JoJo would develop sympathy for his subjects. It is

[149] Maibom (2013) makes a similar point: "For example, our indifference to the death and suffering of nonhuman animals, particularly the ones that we eat, is quite plausibly culpable. We have the capacity to arrive at valuing the life and well-being of nonhuman animals because of values and beliefs that we either posses or that are readily available in our environment. We believe that if an action or policy creates unnecessary or avoidable suffering, then we have a prima facie reason not to perform or institute it. We also know—or if we do not actually know, we could easily come to know—that factory farming creates a great amount of suffering, and that we do not need to consume as much meat as we do for proper nutrition. From those beliefs, there is a relatively straightforward inferential route to the belief that factory farming is wrong and that we ought to oppose policies that permit it" (275).

[150] Faraci and Shoemaker (2010) posit that JoJo is ignorant about the fact that expressions of ill will are morally wrong (328). I think more plausible is that JoJo et al. know that expressions of ill will are in general wrong, but also know that they are sometimes morally justified and confuse these particular cases as cases where ill will is warranted.

[151] It may still be possible that despite the inferential route, JoJo et al. may nevertheless be unable to jump from one moral belief to another. More would need to be said about why this is the case, though for my purposes here, I just need to show that unfortunate moral upbringing alone does not make the jump impossible. Why some agents would be unable to bridge the inferential gap, and how this affects assessments of moral responsibility is an interesting question that warrants further attention.

difficult to imagine that the slaveholder would be completely blinded to his slave's humanity, the Nazi to the Jewish people's plight, the chauvinist to women's strengths and dignity. This is not to suggest that the right values or action should be obvious to these characters, or easy to see and change. What I am suggesting is that in many cases there will be an inferential route from one's existing values and beliefs to morally correct values and beliefs in particular contexts. We know this because *some* people indoctrinated with problematic values and beliefs come to change them at a later point.[152] This entails that the characters above do not lack the *ability* to know right and wrong in particular contexts, though achieving such recognition will often not be easy.[153] Thus, it is instead more accurate to say that, given their upbringings, it is *more difficult* for these characters to see the right action in a particular context. The question, then, is how difficulty in knowing or doing the right thing affects evaluations of praise and blame. Before turning to this question, let me note that we have moved out of the realm of "he acted badly because he was raised that way and couldn't have known otherwise," or "given his upbringing, he lacks the ability to distinguish between right and wrong." People, barring psychological disorder—incurred at birth or through trauma—do have the ability to distinguish between right and wrong, though for circumstantial reasons doing so may be more difficult for some than others. Before I turn to the difficulty question, I want to explore briefly why people who have the ability to acquire correct values and beliefs may fail to do so.

---

[152] Maibom (2013) articulates the point, explaining that if there were no inferential route from current beliefs and values to new ones, we would not see moral progress in the way we have: "*If* it were true that people brought up in societies where slavery was sanctioned by law and common morality were thereby incapable of thinking it was wrong, *then* we should expect no moral change. But such change *did* happen; not in Ancient Greece, but in Europe and the Americas. Therefore, adoption of seriously wrongheaded norms does not, by itself, deprive a subject of responsibility (or reduce said responsibility)" (277-278, emphasis original).

[153] This intuition that JoJo *should* have known better is hypothesized to be driving blameworthy evaluations by participants in a 2010 study by David Faraci and David Shoemaker: "We speculate, however, that what may mitigate the full-blown excusal of JoJo2 is the belief that his ignorance itself is rather culpable, that even though he did not in fact know better (and his ignorance is deep-seated), he *should have*, where this means there were plenty of opportunities for him to infer that expressions of ill will were wrong, if only he had paid closer attention or been sufficiently sensitive" (329).

One reason why people who have the ability to distinguish between right and wrong in a particular context but may not do so is because there is likely vested interest in them maintaining their existing beliefs and values. JoJo, for example, benefits greatly from torturing and imprisoning subjects who express dissent insofar as it promotes fear and thus obedience. Racist Bob—who is presumed white—benefits from his racist beliefs and practices. His beliefs and values tell him that he is a superior being who is entitled to certain kinds of respect and other beings' bodies and property. His beliefs tell him that he is allowed to enact violence on others who threaten these entitlements, or at least treat these others poorly because of their inferiority—e.g. refusing to help them in times of need, profiling them, refusing services or housing to them, refusing to hire them, calling them degrading names, engaging in victim blaming when they are harmed, etc. Bob's racist beliefs allow him to also believe that he 100% deserves the positive things and experiences that he has—that he worked hard for his lot and anyone else could do exactly the same. Other people fail to achieve what he has because of their own lack of work, character, innate qualities, etc. Given these self-serving beliefs, it might be mistaken to think of JoJo, Bob, and the others as innocent children who have been deprived of the appropriate moral education. While it is true that they have been misinformed about the value of others, that misinformation greatly serves their own interests.[154] And this is precisely why this kind of misinformation happens in the first place. Charles Mills writes, for example, in his introduction on white ignorance:

> *White ignorance . . .*
> *It's a big subject. How much time do you have?*
> *It's not enough.*
> *Ignorance is usually thought of as the passive obverse to knowledge,*
> *the darkness retreating before the spread of Enlightenment.*

[154] See also, Charles Mills (2007): "Finally, the dynamic role of *white group interests* needs to be recognized and acknowledged as a central causal factor in generating and sustaining white ignorance" (34).

> *But . . .*
> *Imagine an ignorance that resists.*
> *Imagine an ignorance that fights back.*
> *Imagine an ignorance militant, aggressive, not to be intimidated,*
> *an ignorance that is active, dynamic, that refuses to go quietly—*
> *not at all confined to the illiterate and uneducated but propagated*
> *at the highest levels of the land, indeed presenting itself unblushingly*
> *as* knowledge. (Mills 2007, 13)

Here Mills pushes us to rethink our understanding of some ignorance, to see it as something people may willingly embrace and perpetuate versus have forced upon them.[155] Note that JoJo and racist Bob do not regret that they have to treat people the way they do—JoJo does not fall asleep at night thinking: "ugh, I wish I didn't have to torture people, but alas, it is what it is." If JoJo did think this, we would think he would be able to infer that he *in fact does not* have to torture people. It is part of the hypothetical that JoJo is acting on his genuine values and desires. Michele Moody-Adams also calls attention to the ways in which "affected ignorance" may result in a person who would otherwise be able to distinguish from right and wrong failing to do so (1994, 301). She identifies four types of affected ignorance. In the first, someone uses neutral language to describe abhorrent actions—Moody-Adams gives the example of torturers describing various acts of torture as "the telephone" or "the parrots' swing." Second, people may express a wish to know nothing—Moody-Adams gives an example of an executive who wishes to "know nothing" of the potential wrongdoings of her employees. Third, people may avoid asking questions in a situation that seems to warrant questioning—Moody-Adams gives an example of a mother who repeatedly accepts expensive gifts from her son despite his modest income. Finally, people may avoid acknowledging our human fallibility, that is acknowledging

[155] See also Joan Tronto on the ethic of care: "By this standard, the ethic of care would treat ignoring others—ignorance—as a form of moral evil. We have an unparalleled capacity to know about others in complex modern societies. Yet the temptations to ignore others, to shut others out, and to focus our concerns solely upon ourselves, seem almost irresistible. Attentiveness, simply recognizing the needs of those around us, is a difficult task, and indeed, a moral achievement" (252).

the possibility that either they or others are wrong—Moody-Adams gives two examples here; the first of a bigot who violently silences protest of his bigotry and second of a university administrator who refuses to investigate charges of wronging because his colleague "couldn't possibly" be guilty of sexual harassment[156] (301-302). It seems possible if not likely that something like affected-ignorance #3 or #4 are happening for JoJo et al. They fail to ask questions or imagine that things might be other than they seem. This is surprising given the severity of their actions. One might expect that JoJo lay in bed at night and think about his actions—whether it might be even *possible* that his actions are wrong.[157] Racist Bob might wonder why so many people revere Martin Luther King—who he learned about in his history classes—as a defender of peace and equality. Racist Bob might wonder, if even in a fleeting moment, why over half of voting Americans—many who are white—voted for Obama in 2008 and again in 2012. As we start to think about the details of each of these characters' lives, we will likely find that there *is* evidence that would raise concern about their beliefs and values, and thus the reasons for their wrongdoing are likely more complicated than that this is simply what they were taught. However, my goal here is not to argue definitively that such willful or affected

[156] Mills also gives an example of this last kind of affected-ignorance (though he does not use Moody-Adams' terms) from Herman Melville's *Benito Cereno*: "Boarding a slave ship—the San Dominick, a reference to the Haitian Revolution—which, unknown to the protagonist, Amasa Delano, has been taken over by its human cargo, with the white crew being held hostage, Delano has all around him the evidence for black insurrection, from the terror in the eyes of the nominal white captain, the eponymous Benito Cereno, as his black barber Babo puts the razor to his throat, to the Africans clashing their hatchets ominously in the background. But so unthinkable is the idea that the inferior blacks could have accomplished such a thing that Delano searches for every possible alternative explanation for the seemingly strange behavior of the imprisoned whites, no matter how farfetched" (Mills 2007, 19).

[157] See Joan Tronto: "The failure to be attentive is perhaps most chillingly described in Arendt's account of the 'banality of evil' which she found personified in Adolf Eichmann. Eichmann was unable to focus on anything except his own career and interests; he was simply inattentive and unable to grasp the consequences of what he did except in the most self-centered ways…Arendt has provided an important perspective on evil that we otherwise miss: evil can arise out of ignorance, either willful or established habits of ignorance. If people in the first world fail to notice everyday that the activities spurred by a global capitalist system result in the starvation of thousands, or in sexual slavery in Thailand, are they inattentive? Is this a moral failing? I suggest that, starting from the standpoint of an ethic of care where noticing needs is the first task of humans, this ignorance is a moral failing" (252).

ignorance is happening in the cases of JoJo et al., though it seems possible if not likely.[158] Instead, I aim to interject caution when we imagine these characters we think acted wrongly simply because they were not taught otherwise. Given the prevalence of affected ignorance, I find it more difficult than Wolf to believe that these characters genuinely lack the ability to distinguish between right and wrong.[159]

To recap: the objection initially raised is that my account seems problematic insofar as it holds people responsible for bad values (or a lack of good values) which they could not help but acquire. Or in other words my account of moral responsibility for implicitly acquired and executed values only applies to cases of praise and not cases of blame. I have argued here, using stock cases from the moral psychology literature, that we overestimate the degree to which people's bad values are determined. I have raised significant skepticism about the claim that JoJo et al. lack the ability to distinguish between right and wrong. What does seem plausible, however, is that it is *more difficult*, because of the cultural milieu or explicit teachings, for these characters to recognize their wrongdoing or hold morally appropriate values and beliefs. John, who was explicitly taught by several moral educators that all people deserve equal respect, will

[158] Tronto calls attention to the potential difficulty in parsing out genuine ignorance from willful ignorance or inattentiveness: "But when is ignorance simply ignorance and when is it attentiveness? If I do not know that rain forest destruction happens in order to provide the world with more beef, am I ignorant or inattentive? Suppose that ignorance is built into social structures? Some would argue that one of the consequences of racism, for example, is that Whites do not know, and do not think that they need to know, anything about the lives of Blacks, except for the self-serving myths that they have told themselves" (253).

[159] A question I would have for proponents of the inability thesis at this point is what exactly makes one unable to distinguish between right and wrong in a particular context? Is it simply that they lack exposure to the morally correct idea? For example, no one ever told racist Bob that Black and white people are equal? Is it a more robust indoctrination of a complex of ideas? For example that Bob has been taught a whole web of racist beliefs and ideas which are reinforced by nonracist beliefs and ideas concerning merit and dessert? Because Wolf assumes based on the egregious nature of JoJo's actions that he is unable to distinguish between right and wrong, she does not spend much time unpacking how exactly one arrives in this state. It is also worth noting that common intuitions hold that mere exposure to moral alternatives does not necessarily make one suddenly able to change their beliefs, as shown in a study by David Faraci and David Shoemaker (2010, 327). This should also be confirmed by our everyday experiences, many people who hold morally problematic attitudes have actually been exposed to alternative moral values and belief, but remain unmoved (see also work on confirmation bias). Given such evidence, the bar for being able to change our beliefs seems rather high—surprisingly difficult to meet.

likely have an easier time acquiring good moral values than racist Bob. Thus, the question now at hand is whether difficulty in acquiring the correct moral values should affect our moral evaluations of people's actions.

One might posit here that difficulty in acquiring the correct moral values makes an agent *less blameworthy* for failing to hold the correct moral values or *more praiseworthy* when arriving at the correct moral values despite such difficulty. For example, because it is so difficult for JoJo et al. to see the error in their ways, given their upbringings and cultural milieu, we might still say they are blameworthy for their wrongdoings, but perhaps *less so* than JoJo et al.* who were taught the appropriate moral values, but nevertheless engage in torture, racial hatred, slavery, genocide, etc. Conversely, if it was difficult for egalitarian Erin to acquire morally correct values, due to her upbringing and cultural milieu, we might say that she is *more praiseworthy* for having the right values egalitarian Erin* who was taught the appropriate moral values and acts accordingly. In order to make sense of these claims, we need to revisit scalar concept of blame- and praiseworthiness, meaning the idea that blame- and praiseworthiness exist in degrees. I briefly introduced the idea of "degrees" of praise/blameworthiness above in my discussion about complex moral motivations. Let me say a bit more about the concept and whether it tracks "difficulty" in knowing or doing the right or wrong thing.

Above, I explained that Arpaly holds that degrees of responsibility track depth of concern. When an agent has deeper moral concern, she is more praiseworthy for doing the right thing than someone who does so out of more shallow moral concern (though moral concern nevertheless). For example, imagine that my sister and I volunteer at the local homeless shelter. She does so because she cares deeply about the homeless population and feels passionate about the work. I, on the other hand, care about the homeless population and think the work is

important, but I do not care to the same degree that my sister does. In this case, it seems appropriate to say that we are both morally praiseworthy for volunteering, but perhaps my sister to a higher degree than me. In cases of blameworthiness, we will talk about "lack of moral concern," which may also occur in degrees. Imagine that Sue and David dislike a colleague. Sue simply dislikes the colleague, whereas David has a deep hatred for the colleague. One day, Sue is walking into the building when she notices the colleague walking up behind her. Given that she dislikes him, she is somewhat inclined to scoot in and let the door close before the colleague reaches it. She may waiver for a brief second, because she does have *some* concern for the colleague or doing the right thing, but ultimately that concern is so little that she decides to let the door slam behind her. Later that day, David is in the same situation (poor colleague). He has no concern whatsoever for the colleague or doing the right thing and eagerly lets the door slam behind him. Again, we would likely say here that both Sue and David are blameworthy, but given his total lack of concern, David more so than Sue.

Arpaly begins to unpack "depth of concern" by first explaining what is it not. Depth of concern does not equal intensity of feeling (2003, 85). Being consumed by the desire for a Coke right now does not mean I care more about Coke than my friends. Or feeling distant from my sister after a fight does not mean I care for her any less (85). Depth of concern also does not equal degree of reflective endorsement (85). As shown in the example of my sister and I volunteering, we may equally endorse the volunteer work, but have different degrees of concern. Arpaly brackets giving a full account of concern in general or moral concern in particular as that would "require choosing a theory of desire" (84), but does identify three features/markers associated with the depth of concern, other things being equal (85). The first involves motivation. Arpaly explains that "the more you care about something, the more it would take to

stop you from acting on your concern" (86). If you care deeply about the football team, for example, you show up on game day even when it is raining or 10 below (86). You might go to more games than people who care less. In short, deeper concern often affects one's action. Second, deeper concern results in more complex and intense emotions (86). Recall in Chapter 4 that I explained that I do not care about my houseplants to the same degree that I care about my sister. I may have some emotional responses to my houseplants—pride when they do well, disappointment when they die—but not nearly as complex as the emotions I experience with regard to my sister—joy, grief, fear, love, guilt, anger, etc. Even for common emotions like pride, I feel this emotion more intensely for my sister than my houseplant. In short, depth of concern affects our emotional makeup (86). Finally, depth of concern affects what we notice (86). A person who cares deeply about gender equality will be more likely to notice instances of inequality than a person who doesn't care about gender equality or cares about it to a lesser degree.

Arpaly implies in her discussion about depth of concern that difficulty in doing the morally right thing can help us determine one's depth of concern. She describes someone with "diehard" motivation and her "fair-weathered friend." The person with diehard motivation (the "foul-weathered friend") cares so much for her fellow human beings that she would act benevolently even if severe depression came over her and made it hard for her to pay attention to others. Fair-weathered friend, however, acts benevolently as long as no serious problems distract her; she does good deeds as long as there is no serious crisis in her job or marriage. Finally, contrast these two characters with the friend who acts benevolently on a whim: when she gets a call from a charity asking for a donation, her credit card is sitting next to her and she thinks

"sure, why not?" though would not have donated had her credit card been in the other room. Arpaly concludes:

> The first agent is more praiseworthy for her actions than the second agent, because to act benevolently for moral reasons while one is depressed takes more concern for those moral reasons than to do so in happy times… The third agent—the person whose moral concern is skin deep—may be called the *capricious* philanthropist and would be very presumptuous to expect much praise for an action that almost seems accidental, attributable to the charity's call and the location of the credit card more than the depth of her concern for her fellow human beings. (2003, 87-88)

In short, depth of concern should affect whether one would do the right thing in difficult situations.

However, I suspect that many infer from analysis like Arpaly's that doing the right thing in difficult situations indicates deeper concern than those who do the right thing in easier situations. In other words, if depth of concern leads to right action in difficult situations, we might think that acting in difficult situations indicates deeper concern. Because it is more difficult for the depressed person to act benevolently, we infer that she has deeper moral concern that the friend who acts benevolently, but in times when it is easier to do so.[160] This confirms the hypothesis that someone is more praiseworthy when they do something right under difficult circumstances and less blameworthy when they fail to do the right thing under difficult circumstances. This also coheres with common intuitions. In a study, David Faraci and David Shoemaker show that participants rate JoJo as morally blameworthy (on average 4.77 out of 7, where 7 is completely blameworthy, 4 is somewhat blameworthy, and 1 is not at all blameworthy).[161] This shows that participants do in fact find JoJo blameworthy, though perhaps evaluate him less harshly because it is more difficult for him to do the right thing given his upbringing than it is for his father, who seems to have more explicitly chosen his values and

[160] Arpaly does not argue for this idea explicitly, though I could see how one might infer it from her analysis.
[161] In a more recent study, the character representing JoJo scores 5.4/7. Faraci and Shoemaker 2014.

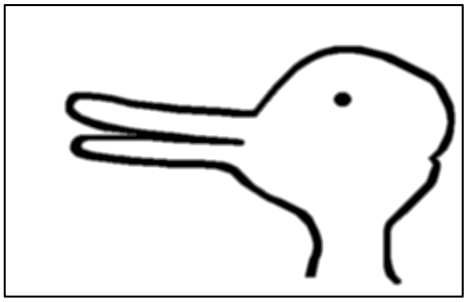
beliefs (Jo Sr. is ranked on average 5.8 out of 7 on the blameworthy scale) (2010). Faraci and Shoemaker call this the difficultly hypothesis: "actions are more or less attributable to agents in these sorts of cases depending on the degree of difficulty they are judged to have in recognizing various features of their actions about which they remain ignorant" (2010, 331. See also Faraci and Shoemaker 2014).

However, while it might be common intuition to hold actions performed in difficult situations more or less praise/blameworthy, there is reason to doubt that difficulty should affect our evaluations of praise and blame. Let us revisit the volunteer example with my sister and I, but this time let us presume that we both care deeply about the cause. Further presume that it is very easy for her to volunteer—she is young, single, healthy, and has a very flexible schedule. Imagine that, I, on other hand, have health problems, familial obligations, and very little free time. It is thus much more difficult for me to volunteer than it is for her. This, however, does not mean that I deserve more praise than she does. She cares about the cause just as much as I do, if not more. Thus, in this case, the degree of praise does not track onto difficulty. Let us imagine another case. Imagine that Jon has been a social justice advocate his entire life. He grew up with very socially active parents and continues to be passionately engaged in the work. Let us further imagine that George was raised by very apathetic and negative parents, which he had absorbed and followed until college, where he was slowly influenced through classes and friends to pursue social justice work, which he has found important and fulfilling much to his surprise. John and George have been working together on a campaign to end sexual harassment on campus—both care deeply about the cause and work tirelessly. Does the fact that it was more difficult for George to arrive at his current values and beliefs mean that he is more praiseworthy than Jon? It would be odd, and perhaps undercutting of Jon, to say so. Now let us turn to blameworthy cases.

Sally and Kristen have decided to engage in identity theft to make some extra money. Sally has been hacking for a long time and can easily access people's information. Kristen, on the other hand, has no experience and has to work very hard and practice great patience to access people's data. It is frustrating and she often feels like giving up but resolves to stick with it. Is Sally more blameworthy than Kristen? Less? JoJo's cousin, KoKo, has also been raised to accept his father's (also an evil dictator) values and belief. KoKo, for some inexplicable reason, however, is appalled about torturing and killing, though nevertheless does it, while JoJo fully embraces the practice. JoJo and KoKo engage in the same actions and both were raised in the same circumstances that made it difficult to see the right thing to do, are they equally morally blameworthy? That the answers, or our intuitions push in the other direction or are not clear in these cases indicates that difficulty does not so easily and directly determine degrees of praise/blameworthiness.[162] Why do our intuitions seem so strong in other cases, then?

I suspect that we cut JoJo some slack in comparison to Jo Sr. simply because we are impressed when people overcome difficult circumstances or challenges, and not surprised when they don't. This does not necessarily mean they are more praiseworthy or less blameworthy for their actions than their counterparts who do not face difficult circumstances and express the same actions. It just means that they do something impressive when they do the right thing in the face of challenging circumstances or that they are more normal or average when they fail to do the right thing in the face of challenging circumstances. Faraci and Shoemaker (2014) give a useful analogy for this point. They ask us to imagine that someone has been shown the famous image of the "duckrabbit" repeatedly since childhood:

[162] See Markovits 2010 and Nelkin 2014, for example, about skepticism of the role that difficulty may play in determining degrees of responsibility.



Further imagine that the person has been taught over and over that what he is looking at is a duck and only a duck. When he meets you as an adult and you keep insisting that the image is also a rabbit, it will be no surprise that he has a hard time seeing it so. But now imagine that his sister, raised with the same teachings, enters the room and, upon being told that the duck is also a rabbit, sees the rabbit immediately. Faraci and Shoemaker think that we will probably applaud her ability to see past what she is used to. Relatedly, we probably won't berate the brother and might even understand why he does not easily see the rabbit, but we might also have a sense that really he should be able to, even though doing so might be difficult for him. The idea, then, is that it is impressive when people exceed expectations. The same intuitions are likely influencing our judgments of JoJo and Huck: JoJo meets our expectations while Huck surpasses them. Note though that overcoming the difficulty in itself may not be morally relevant. Nelkin (2014) makes the point:

> Perhaps what explains the appearance that difficulty is related to responsibility is that difficulty *is* related to *non*-moral praiseworthiness. We praise people who accomplish difficult tasks, whether climbing mountains or solving math problems. Difficulty of the task yields greater praiseworthiness. Perhaps, then, in the cases in which someone acts well, with great effort, there is simply a fixed amount of moral praiseworthiness that does not track difficulty, combined with a variable amount of non-moral praiseworthiness that is not different in kind from that which one deserves for, say, reaching the top of Machu Picchu. (14)

Returning back to JoJo, then, we may not be surprised that he doesn't overcome his difficult circumstances, but also recognize that he does immoral things because of a lack of concern for others (which he should have). Huck on the other hand, does the right thing out of deep moral concern that we are surprised, or impressed, that he has developed. In short, I am not convinced that difficulty actually affects the degree of moral responsibility, though I see why it pushes our intuitions in that direction.

I suspect that what we actually care about, as Arpaly suggests, is the depth of someone's concern and difficulty stands in as a heuristic for depth of concern. As I talked about in Chapter 3, we want people to do kind things for others and us because they care, not in a superficial way, but that in a robust and deep way. I posit that we often use difficulty as a heuristic for depth of care. This seems to be going on in Arpaly's analysis; she takes the fact that the depressed person acts benevolently in spite of her depression as indicative of the fact that her concern for others runs deep. Similarly, we might guess that the person who acts benevolently in easier times might not have such a depth of care. But note from my examples that this inference is where we get into trouble—someone can have great depth of concern *and* do good in easier times. So while difficulty might often be a good heuristic for depth of concern, it is not 100% reliable. In fact, I think this explains why our intuition is to cut JoJo some slack—we think that JoJo has some degree of concern, but because the difficulty in doing the right thing in his situation is so high, and the degree of concern is somewhat minimal, he nevertheless does the wrong thing. In different circumstances, JoJo might instead do the right thing. This cannot be said for Jo Sr. We get the impression that even in different circumstances, Jo Sr. would act immorally. This indicates a complete lack of concern for others. Thus, we find him especially blameworthy given this total lack of concern. We aren't sure exactly what the depth of concern is for JoJo, but do

have the sense that if placed in different circumstances, he might be able to express it. My main point here is that it isn't difficulty per se that affects degrees of responsibility, but rather the depth of one's concern. We often use difficulty and ease to determine the depth of one's concern, but this is merely a heuristic and should be used with caution.

Additionally, we might be inclined to cut JoJo et al. slack because we pity those for whom it is difficult to see or do the right thing. Feeling such pity seems morally permissible and not necessarily in tension with saying that someone is blameworthy. We likely imagine that JoJo is missing out on meaningful relationships and the deep joy of helping others. Even though he is in his father's favor, it is probably a tenuous relationship. JoJo might genuinely feel his father's affection for him, but might also be afraid of his father given his brutal and arbitrary treatment of others. We might pity racist Bob for the hatred that he harbors and, like JoJo, the loss of potentially meaningful relationships and opportunity to find joy in making the world more equitable. These characters do not live enviable lives. Of course, that in itself does not mean they are not blameworthy for their malevolent actions. It simply means we have complicated feelings about these characters that might influence our moral evaluations of them. Again, let us look at Walter White. Walt commits the most horrific acts, which could in no possible way be justified, and does so as a fully rational and deliberative agent. Nevertheless, we cannot help but feel bad for Walt as he struggles with his cancer and is banned from seeing his children. It is not incoherent to blame and pity morally bad actors.

Finally, I suspect that we are more hesitant to blame someone for something that appears out of their control, while willing to praise someone for something that seems out of their control because the stakes and praise and blame are so different in practice. If we incorrectly blame someone, we have acted unfairly, perhaps mistreated someone. However, if we incorrectly praise

someone, we have not really caused them harm at all. So, we may be particularly vigilant when assigning evaluations of blame in comparison to evaluations of praise (thus the seeming asymmetry raised at the beginning of this section). In her discussion of asymmetrical freedom, Wolf articulates such a point:

> ….I think, we have a stronger reasons for wanting acts of blame [versus acts of praise] to be justified. If we blame someone or punish him, we are likely to be causing him some pain. But if we praise someone or reward him, we will probably only add to his pleasures. To blame someone undeservedly is, in any case, to do him an injustice. Whereas to praise someone undeservedly is apt to be just a harmless mistake. (155-156)

I suspect that this asymmetry in the consequences of praise and blame in part drives our intuitions that while it might be morally permissible to praise someone for her implicitly acquired and executed morally good values, it is impermissible to blame someone for his implicitly acquired and executed bad values. However, the fact that there seems to be higher stakes in incorrectly blaming does not mean that someone like JoJo is not blameworthy, but rather that we might need to exercise extra caution in assigning evaluations of blame.[163] It might also be useful to recall the distinction here that I made in Chapter 3 between evaluations and practices of blame and praise. In my project here, I have been concerned solely with the standards and conditions of evaluations of responsibility, praise, and blame. What the standards and conditions of practices of praise and blame ought to be is a separate question, though important. We should keep this distinction in mind in the cases I have discussed here because it may be possible that JoJo et al. are in fact blameworthy, though for pragmatic reasons or other considerations we might forgo, or mitigate, the practice of blaming them. This is a question worthy of further research.

[163] Though I think there are also harmful consequences when we suspend judgments of blame in cases where we do not have a strong degree of certainty regarding the situation or the agent's intentions. For example, in many cases of sexual assault, we will lack definitive evidence of wrongdoing, but we may create more harm for the victim and perhaps society more broadly by suspending evaluations of responsibility given the lack of definitive evidence.

To recap this section: I presented the concern that my account of value-guided automaticity, according to which we are morally responsible for implicitly acquired and executed values, only applies to cases where an agent has acted in a morally praiseworthy way. The concern emphasizes that it would be unfair to hold people responsible for bad values, or lack of morally good values, that they could not help but have. I then explored what exactly we mean when we say that an agent "could not help" but have the values and beliefs he does. Using cases from Susan Wolf's "Sanity and the Metaphysics of Responsibility" I explored several possible hypotheses: that JoJo et al. lack the *capacity* to distinguish between right and wrong, that they lack the *ability* to distinguish between right and wrong full stop, that they lack the ability to distinguish between right and wrong in the relevant *particular context*, and that it is *difficult* for them to see and do the right thing. I raised reasons to doubt that these hypotheses properly explain why JoJo et al. engage in wrongdoing and fail to exculpate them from moral blameworthiness.

Conclusion and Future Research

I started this dissertation with Haidt's claim that we rarely engage in conscious deliberation about moral issues and actions, and when we do it is often only to justify the intuitive judgments at which we have already arrived (Chapter 1). Haidt takes himself to be dealing a major blow to many philosophers working in moral psychology. In Chapter 2, I discuss descriptive philosophical responses to Haidt's work, ranging from arguments that we do in fact deliberate about moral issues, contra Haidt's data, to the explanation that judgments that may automatically and subconsciously be triggered are nevertheless deliberatively developed. I argued that these kinds of responses miss the point of Haidt's analysis and the opportunity to further explore

unconscious cognitive processing. I posited that philosophers miss this opportunity in part because of deeply entrenched biases about the importance of reflection, but also because of Haidt's divisive tone and over-simplification of the automatic processes underlying moral judgment and decision-making. I argued that despite significant, not entirely unwarranted, skepticism about Haidt's work, there has been robust research in the last several decades on the pervasiveness and influence of automaticity in various arenas of moral judgment and decision-making. I specifically discussed data that suggest that automatic processing can do a better job in some contexts, measured in games of cooperation for example, of leading us to the morally good action or outcome. And I talked about cases in which conscious deliberation makes us perform worse—such as in sports or improvised music. I concluded the following in Chapter 2:

- *There are two different processes/representations/systems that guide judgments, decisions, and behaviors*.
- *Both of these processes make important contributions in guiding our judgments, decisions, and behaviors.*
- *It appears that both processes are somewhat situationally constrained—meaning they function well in some contexts or situations, but not others.*
- *Both processes need and deserve further study.*

I then considered, in Chapter 3, normative responses to Haidt's work, which held that even if Haidt's work was empirically adequate, it does not affect our normative theories of good moral judgment and decision-making. I explored various reasons for why we might think that even if deliberation is not playing a leading role in much of our moral judgment and decision-making, it *should*—for example that actions which we do not deliberate about are not genuinely *ours* (not expressions of our agency) or that they are merely accidentally right or wrong. I argued, however, following the philosophical methodology of theorists like George Sher, Nomy Arpaly, and Angela Smith, that there are many cases in which someone does not act from conscious deliberation, but is nevertheless evaluated as praiseworthy. Folks like Wesley Autrey, Holocaust

rescuers, and Huckleberry Finn performed morally admirable actions without engaging in deliberation about the action in the moment or at previous junctures. I then explained, drawing upon the work of Arpaly, Smith, and Julia Markovits that what matters for moral responsibility (and in turn praise and blame) is not whether one deliberates but whether one acts for the right reasons (or in cases of blame for the wrong reasons, or lack of right reasons). In Chapter 4, I presented my own account of how good moral judgments and decisions can be guided exclusively by the automatic system and yet still be robust enough to establish moral responsibility. I argued that automatic processing engages with implicitly held moral values, which in turn entail the right reasons for acting discussed in Chapter 3. I see my account as taking seriously Haidt's call to better understand automaticity, while more satisfactorily responding to philosophical concerns about the lack of normative content in automatic processes. And here in Chapter 5, I have explained how my account not only handles cases of straightforward praiseworthiness, but also more complicated cases where individuals have conflicting moral values or good values that go bad, as well as cases where individuals have implicitly acquired and held bad moral values. My aim has been to convince readers that, as Haidt encourages, we ought to be pursuing more research on automaticity and moral judgment and decision-making. There is reason to believe that automaticity guides *much* of our moral judgment and decision-making. This is not a fact to fear, but to embrace and aim to further understand—not because automaticity must be controlled, but because it might have greater potential for moral intelligence and success than we currently appreciate.

In closing, I would like to identify a few potential areas and questions for such future research. First, there is more work to be done to gain better understanding of when and how automaticity works well. In what kinds of cognitive and environmental contexts does it foster

good moral judgment and decision-making and in what kinds of contexts does it function poorly? For my account specifically, there is much more empirical research needed on values: what values are, how exactly we acquire them both implicitly and explicitly, how exactly they guide our moral judgments and decisions. I argued in Chapter 4 that we have good reason to think that values can be acquired, triggered, and executed automatically, but there is still much more research to be done to confirm and explain this phenomenon. Another interesting line of research would look at whether there is a difference in value-acquisition and expression across social groups in the United States as well as cross-culturally. These areas of research will involve questions from a variety of disciplines, including cognitive science, psychology, pedagogical studies, and perhaps sociology. In more philosophical or normative contexts, there is more work to be done in my account on the connection between care and morality. I also plan to articulate in future research more precisely how my account is feminist. And finally, there is much work to be done on the implications of an account like my value-guided automaticity for theories of moral education.

Works Cited

Applebaum, Barbara. 1997. "Good Liberal Intentions Are Not Enough! Racism, Intentions and Moral Responsibility." *Journal of Moral Education* 26(4): 409-21.

Arendt, Hannah. 1963. "Eichmann in Jerusalem—Il" *The New Yorker* February 16, 1963 Issue.

Bedrick, Jeffrey. 2014. "The "Reasonable Person" and the Psychopath." *Philosophy, Psychiatry, & Psychology* 21(1): 13-15.

Blair, R.J.R., Blair, L., Jones, F., Smith, M., and Clark, M. 1995. "Is the Psychopath 'morally Insane'?" *Personality and Individual Differences* 19(5): 741-52.

Blair, Rjr. "Responding to the Emotions of Others: Dissociating Forms of Empathy through the Study of Typical and Psychiatric Populations." *Consciousness And Cognition* 14, no. 4 (2005): 698-718.

Buss, Sarah. 1997. "Justified Wrongdoing." *Noûs* 31(3): 337-69.

Ciurria, Michelle. 2014. "The Case of Jojo and Our Pretheoretical Intuitions: An Externalist Interpretation." *Review of Philosophy and Psychology* 5(2): 265.

Faraci, David and David Shoemaker. 2014. "Huck vs. JoJo: Moral Ignorance and the (A)symmetry of Praise and Blame." In *Oxford Studies in Experimental Philosophy*, Oxford Studies in Experimental Philosophy, Chapter 2. Oxford University Press.

Faraci, and Shoemaker. 2010 "Insanity, Deep Selves, and Moral Responsibility: The Case of JoJo." *Review of Philosophy and Psychology* 1(3): 319-32.

Greenspan, Patricia S. 2003. "Responsible Psychopaths." *Philosophical Psychology* 16(3): 417-29.

Law, Iain. 2003. "Autonomy, Sanity and Moral Theory." *Res Publica* 9(1): 39-56v

Levy, Neil. 2003. "Cultural Membership and Moral Responsibility." *The Monist* 86(2): 145-63.

Levy, Neil. 2005. "The Good, the Bad and the Blameworthy." *Journal of Ethics & Social Philosophy*, 1(2): 1-16.

Mason, Elinor. 2015. "Moral Ignorance and Blameworthiness." *Philosophical Studies* 172(11): 3037-057.

Maibom, Heidi L. 2013. "Values, Sanity, and Responsibility." In *Oxford Studies in Agency and Responsibility Volume 1*, Oxford Studies in Agency and Responsibility Volume 1, Chapter 12. Oxford University Press.

Maibom, Heidi. 2008. "The Mad, the Bad, and the Psychopath." *Neuroethics* 1(3): 167-84.

Markovits, Julia. 2010. "Acting for the Right Reasons." *Philosophical Review* 119(2): 201-42.

Mills, Charles W. 2007. "White Ignorance." In *Race and Epistemologies of Ignorance*, 13-38. Eds Shannon Sullivan and Nancy Tuana. Albany: State University of New York Press.

Moody-Adams, Michele M. 1994. "Culture, Responsibility, and Affected Ignorance." *Ethics* 104(2): 291-309.

Narváez, Darcia. 2014. *Neurobiology and the Development of Human Morality: Evolution, Culture, and Wisdom*. Norton Series on Interpersonal Neurobiology.

Nelkin, Dana Kay. 2015. "Psychopaths, Incorrigible Racists, and the Faces of Responsibility *." *Ethics* 125(2): 357-90.

Nelkin, Dana Kay. 2014. "Difficulty and Degrees of Moral Praiseworthiness and Blameworthiness." *Noûs.* Published on Wiley Blackwell Early View 2014. DOI: 10.1111/nous.12079

Scanlon, T. M. 2012."Interpreting Blame." In *Blame*, Blame, Chapter 5. Oxford University Press.

Shoemaker, David. 2003. Caring, Identification, and Agency. *Ethics, 114*(1), 88-118.
Shoemaker, David. 2011. "Psychopathy, responsibility, and the moral/conventional distinction." *The Southern Journal of Philosophy* 49: 99-124.
Shoemaker, David. 2015. *Responsibility from the margins*. Oxford: Oxford University Press
Sripada, Chandra. (2015). Self-expression: a deep self theory of moral responsibility. *Philosophical Studies* Published online August 12, 2015 doi 10.1007/s11098-015-0527-9
Talbert, Matthew. 2008. "Blame and responsiveness to moral reasons: are psychopaths blameworthy?" *Pacific Philosophical Quarterly* 89(4): 516-35.
Talbert, Matthew. 2012. "Moral Competence, Moral Blame, and Protest." *The Journal of Ethics* 16(1): 89-109.
Tronto, Joan. 2005. *An Ethic of Care*. In Cudd, Ann E., and Andreasen, Robin O (Eds) *Feminist Theory : A Philosophical Anthology*. 1st ed. Blackwell Philosophy Anthologies; 23. Oxford, UK; Malden, MA: Blackwell Pub.
Vincent, Nicole A. 2011. "Madness, Badness, and Neuroimaging-Based Responsibility Assessments." In *Law and Neuroscience*, Law and Neuroscience, Chapter 6. Oxford University Press.
Vogel, Lawrence. 1993. "Understanding and Blaming: Problems in the Attribution of Moral Responsibility." *Philosophy and Phenomenological Research* 53(1): 129-42.
Watson, Gary. 2011. "The Trouble with Psychopaths." In *Reasons and Recognition*, Reasons and Recognition, Chapter 13. Oxford University Press.
Watson, Gary. 2013 "XIV—Psychopathic Agency and Prudential Deficits." *Proceedings of the Aristotelian Society (Hardback)* 113(3.3): 269-92.
Wolf, Susan. 1980. "Asymmetrical Freedom." *The Journal of Philosophy* 77(3): 151-66.
Wolf, Susan. 1987. "Sanity and the Metaphysics of Responsibility." In Ferdinand David Schoeman (ed.), *Responsibility, character, and the emotions: New essays in moral psychology*. Cambridge [Cambridgeshire]; New York: Cambridge University Press. 46-62.