Revista de Humanidades de Valparaíso, 2022, No 19, 85-106 *Humanities Journal of Valparaíso*
DOI: https://doi.org/10.22370/rhv2022iss19pp85-106 eISSN 0719-4242
Sección Monográfica / Monographic Section

Teaching an Old Dog New Tricks: Intuition, Reason, and Responsibility

Enseñándole trucos nuevos a un perro viejo: intuición, razón y responsabilidad

Stephen Setman
Purdue University, United States of America
stephensetman@gmail.com

Abstract

According to one highly influential approach to moral responsibility, human beings are responsible (eligible to be praised or blamed) for what they do because they are *responsive to reasons* (Fischer & Ravizza 1998). However, this amounts to a descriptive assumption about human beings that may not be borne out by the empirical research. According to a recent trend in moral psychology (Haidt 2001), most human judgment is caused by fast, nonconscious, and intuitive processes, rather than explicit, conscious deliberation about one's reasons. And when humans do engage in explicit deliberation, it primarily serves to provide post hoc rationalization of their intuitive judgments (confabulation). If this is correct, it is tempting to conclude that most of our judgments—and the actions we perform on their basis—are not genuine responses to reasons. The reasons-responsiveness approach would thus appear to be committed to the implausible conclusion that we are not responsible for very much after all, including, most problematically, our implicit biases. I argue that the reasons-responsiveness approach can avoid this conclusion by showing three things: (1) that affective and intuitive processes can be reasons-responsive; (2) that the responsiveness of those processes can be bolstered by the agent's environment; and (3) that practices like blame are one of the key ways in which human beings are attuned to reasons over time. I argue that the first and second of these items, despite their initial plausibility, are insufficient on their own to explain why humans can be held accountable for things like implicit biases, and that the way forward is to appreciate what holding each other accountable *does*—i.e., its effects.

Received: 02/12/2020. Final version: 27/03/2022
eISSN 0719-4242 – © 2022 Instituto de Filosofía, Universidad de Valparaíso

Keywords: responsibility, blame, blameworthiness, intuition, reason, reasons-responsiveness, control, ecological control, moral ecology, scaffolding, moral psychology, dual process, Jonathan Haidt.

Resumen

Según una teoría influyente de la responsabilidad moral, los seres humanos son responsables (pueden ser disculpados o culpabilizados) cuando tienen la capacidad de *responder a razones* (Fischer & Ravizza 1998). Pero esta teoría hace una suposición descriptiva sobre los seres humanos que posiblemente no es consistente con investigaciones empíricas. De acuerdo con una tendencia reciente en la psicología moral (Haidt 2001), la mayor parte del juicio humano es resultado de un proceso inconsciente, rápido, e intuitivo en vez de un proceso consciente, deliberado, y racional. Y cuando el ser humano participa en la deliberación consciente, generalmente esta sirve para dar una justificación *post hoc* de sus intuiciones ("confabulación"). Si este modelo es correcto, uno puede concluir que la mayoría de nuestros juicios no están basados en una respuesta genuina basada en razones. Así, esta influyente teoría de la responsabilidad moral parece llevarnos a la conclusión de que no somos responsables por muchos de nuestros pensamientos y comportamientos, incluyendo los prejuicios implícitos. En este trabajo discuto que esta conclusión puede ser evitada mostrando: (1) que nuestras inclinaciones y procesos intuitivos pueden *responder a razones*; (2) que estos procesos pueden ser fortalecidos por el ambiente del individuo; y (3) que las prácticas sociales, como la culpa, son una forma clave en que los humanos se ajustan a las razones. Argumento que los puntos primero y segundo, a pesar de su plausibilidad inicial, son insuficientes para explicar por qué los humanos somos responsables de cosas tales como nuestros prejuicios implícitos, y que el modo de progresar en relación con este problema radica en apreciar los *efectos* de las prácticas sociales tales como la culpa.

Palabras clave: responsabilidad, culpa, culpabilidad, intuición, razón, sensible a razones, control, control ecológico, ecología moral, andamiaje, psicología moral, proceso dual, Jonathan Haidt.

1. Introduction

According to one highly influential approach to moral responsibility, human beings are responsible for what they do only if they are *responsive to reasons* (Fischer & Ravizza 1998; Smith 2003). For example, an agent is blameworthy just in case three conditions are met (McGeer & Pettit 2015): First, she had the capacity to recognize and respond to the reasons in her situation. Second, the agent failed to exercise this capacity. Lastly, her failure is not explained by an excusing factor.

The reasons-responsiveness approach thus makes a descriptive assumption about human psychology that may not be borne out by the empirical research. According to a popular "dual process" model of human psychology (Wason & Evans 1975; Frankish 2010), most human judgments, including moral judgments (Haidt 2001), are caused by fast, nonconscious, and intuitive processes, rather than explicit, conscious deliberation about one's reasons. And when humans do engage in explicit deliberation, it primarily serves to provide post hoc rationalization of their intuitive judgments (confabulation). If this is correct, it is tempting to conclude that most of our judgments—and the actions we perform on their basis—are not genuine responses to reasons. The reasons-responsiveness approach would thus appear to be on shaky ground. Either it is committed to the implausible conclusion that we are not responsible for much of what we think and do, including especially our implicit biases, or it will need to give an account of reasons-responsiveness which can explain human beings' responsibility for such things. The goal of this paper is to show that the latter option is a viable one.

As I'll discuss, the most obvious first move for the reasons-responsiveness theorist would be to argue that affective and intuitive processes can be reasons-responsive—and this idea does appear to be supported by recent developments in affective neuroscience. But this can only be a partial solution, because it cannot explain why human beings are sometimes blameworthy when their intuitive processes *fail* to respond to reasons. Even if human intuition is capable of being attuned to the right kinds of reasons, so long as that attunement is contingent on fortuitous circumstances which the agent is not responsible for (e.g., what Joshua Greene (2017) calls "good data" and "good training"), it remains unclear how the agent could be blameworthy for her moral failures.

Some theorists have made promising advances on this problem by appealing to Andy Clark's (2007) notion of *ecological control* (Holroyd & Kelly 2016; Washington & Kelly 2016). However, once this account is made more precise, it becomes clear that an agent's failure to exercise ecological control may still be traceable to factors for which she is not responsible. This objection turns on a general problem with backwards-looking approaches to blameworthiness, which Victoria McGeer and Philip Pettit (2015) have dubbed the "Hard Problem" of responsibility. Although they do not present the Hard Problem in the context of implicit biases and other intuitive processes, I believe the problem and their solution to it provides a way forward for the reasons-responsiveness approach. That solution works by enriching our understanding of an agent's "moral ecology" (Vargas 2013), so that it includes the very practices whose justifiability is in question. That is to say, because praise and blame partly enable the agent's capacity to respond to reasons, these practices can be justified by their forward-looking effects. On this picture, praise and blame are very sort of "good data" and "good training" on which our capacity to recognize and respond to reasons—including reasons of a specifically social or moral variety—depends.

After sketching the reasons-responsiveness approach (Section 1) and the challenge presented by dual process theories (Section 2), I will consider two steps in the right direction

which theorists have already taken—Peter Railton's (2014; 2017) argument that the affective system can be reasons-responsive (Section 3), and Natalia Washington and Daniel Kelly's (2016) and Jules Holroyd and Daniel Kelly's (2016) arguments that responsibility is partly grounded in an agent's environment (Section 4). Along the way I explain why each step falls short of an adequate reasons-responsiveness account. I call them steps in the right direction, because the account I ultimately defend recruits and builds upon them both. I argue that these advancements, if combined with the forward-looking notion of reasons-responsiveness defended by McGeer and Pettit (2015), can explain why human beings are sometimes blameworthy even for things which result from unconscious, intuitive processes (Section 5). Then (Section 6) I elaborate on the resulting view and provide two clarifications. The old "emotional dog" that we inherited from evolution and fortuitous learning environments may not be under any single agent's direct, conscious control, but it is capable of learning new tricks.[1] And it is capable of learning precisely because we agents hold one other accountable for our reasons-responsiveness failures.

2. The Reasons-Responsiveness Approach

The central claim of the reasons-responsiveness approach is that an agent is responsible for something, such that she may be praised or blamed for it, only if she acted from her own, "reasons-responsive" process (Fischer and Ravizza 1998, 38-39).[2] We can think of reasons here as any consideration, usually some feature of the situation at hand, which speaks for or against various courses of action or judgments. They could be merely prudential reasons concerning the means of achieving desired ends, or they could be normative reasons concerning what is impermissible, permissible, or required of the agent. Importantly, to say that an agent acts from her own, reasons-responsive process is to say that the agent is capable of recognizing those aspects of her situation which really do carry practical or normative significance, and that she is capable of choosing and acting as they require, permit, or support. If an agent fails to respond to her situation appropriately, say by violating a legitimate norm that prohibits the behavior in question, then she is blameworthy only if (1) she had the capacity to respond to the reasons in her situation, (2) she failed to exercise this capacity, and (3) her failure is not explained by an excuse.

Reasons-responsiveness was originally presented by John Martin Fischer and Mark Ravizza (1998) as a way of filling out the Aristotelian "control condition" of responsibility that did not rely upon the problematic notion of alternative possibilities. The idea that control is necessary for responsibility is not universally accepted, but the basic idea is that, to justifiably

[1] This pun turns on the title of Jonathan Haidt's landmark essay, "The Emotional Dog and Its Rational Tail" (2001), and the English idiom, "You can't teach an old dog new tricks," which is supposed to suggest that it is very difficult to change people's habits.

[2] Fischer and Ravizza primarily speak in terms of "mechanisms" of action, but say that "we could instead talk about the process that leads to the action".

praise or blame someone for something, that thing must have been "up to her" in some sense. On early compatibilist approaches, this just meant that the agent could have done something else if she had wanted to—she was not compelled or forced to act as she did. But as Harry Frankfurt (1969) argued, this interpretation of control cannot explain why agents are still intuitively responsible for what they do, even if they couldn't have done otherwise due to a "counterfactual intervener". In light of Frankfurt's challenge, Fischer and Ravizza argued that, if we focus instead on the actual process which led to the agent's action and consider whether that process was responsive to reasons, we can explain this intuition. Even if the agent could not have actually done otherwise, she was still the one guiding her behavior through her own recognition and response to reasons, and she was still capable of recognizing the reasons there were in her situation and responding to them appropriately. In their view, this "guidance control" is all the control required for responsibility.

Fischer and Ravizza identify two separate dimensions of guidance control, corresponding to two conditions. One is that the process leading to the agent's action must be the "agent's own" (1998, 170-202). So, the action under consideration must have resulted from the *agent's own* mental process. By contrast, if an action resulted from a causal process which does not belong to the agent, such as a device which, unbeknownst to the agent, had been implanted in her brain, then the process which resulted in that action is not really the agent's, even if it took place within her body. Besides clandestine brain implants, there are other ways in which a process can fail to belong to the agent. Fischer and Ravizza mention hypnosis, brainwashing, and subliminal advertising as ways an agent could come to act from processes which are not her own, due to the history behind these processes (1998, 196-197).

The second condition concerns the degree to which the agent's process is responsive to reasons (Fischer and Ravizza 1998, 41-46). Recall that this notion was introduced as a way of making sense of the control requirement for responsibility without appealing to alternative possibilities. In order for an agent to be justifiably blamed for her failures, it would seem that the agent must have been capable of acting in the way she is putatively required to act. Blame implies that the agent has done something she *should not have done.* So, implicit in blame is a claim to the effect that this agent *should have done otherwise.* For example, suppose Mariem should have yielded to let a group of pedestrians cross the road, but failed to do so.[3] If ought implies can, then Mariem must also have been capable of yielding, since otherwise we would be demanding something of Mariem which she could not have done. According to the reasons-responsiveness approach, the sense in which Mariem "could have done otherwise" is that she acted from a process which is reliably sensitive to reasons, understood as a modal property of the process. If Mariem, utilizing the same kind of process, would have yielded to

[3] Moreover, when we blame Mariem, we do not just imply that she should have yielded. We are also claiming that she should have done so *for certain reasons*—the safety of the pedestrians, their right-of-way, etc. Her failure is not just a failure to execute certain bodily movements, but is a failure to recognize and respond to *the reasons there were* to execute those movements.

the pedestrians in a suitable range of similar, hypothetical scenarios (or at a suitable number of possible worlds), then we can say that Mariem was capable of acting as she should have in the actual scenario.[4] However, if the same process would be unresponsive to relevant changes in the situation, exhibiting a kind of inflexibility or compulsiveness, for example, then this tells us something about the way Mariem acts in the actual scenario—namely, from a process that is not sufficiently responsive to reasons.[5]

Although Fischer and Ravizza do allow that a process may be reasons-responsive even if it does not involve conscious deliberation about one's reasons (1998, 85-89), the way they most commonly characterize responsiveness to reasons uses a "perceive-think-act" model of human agency. On this model, responsiveness to reasons is mediated by a process of practical reasoning in which the agent consciously considers and weighs the reasons in her situation. In the cases they present Fischer and Ravizza regularly make appeals to the agents "normal faculty of practical reasoning" as a process "which we can reasonably take to be reasons-responsive" (1998, 38), and they describe the process of responding to reasons as (1) *taking* reasons to be sufficient (which they call receptivity to reasons), (2) *choosing* in accordance with those reasons (reactivity to reasons), and, finally, (3) *acting* in accordance with the choice (Fischer and Ravizza 1998, 41). Arguably, taking reasons to be sufficient for some course of action and choosing in accordance with that judgment are most naturally understood as consciously mediated processes of practical reasoning.

With this exposition in the foreground, consider again the challenge raised by dual process theories like Haidt's social intuitionism. If the "normal human faculty of practical reasoning" (Fischer and Ravizza 1998, 42) is meant to serve as a paradigmatic example of a reasons-responsive process, and if most human judgment is actually caused by nonconscious, intuitive processes, rather than explicit deliberation, it is tempting to conclude that most of our judgments (and the actions we perform on their basis) are not genuinely responsive to reasons. It will at least be necessary to show how processes that do not involve the explicit consideration and weighing of reasons can nonetheless be responsive to reasons in the sort of way required by condition two above. Otherwise, although human beings *think* they do what they do (and judge what they judge) for reasons, and even though they are eager to provide reasons for their actions and judgments upon request, it may turn out that they are simply mistaken. Most of the time they are "reasons-blind." The reasons-responsiveness theorist thus faces a choice. She must either accept that human beings are not responsible for the majority of what they do (including, most problematically, the things they do on the basis of

[4] On their view, responsiveness to reasons can be *strong* or *weak,* depending on the range of possible worlds at which this is true. See Fischer & Ravizza (1998, 41-46).

[5] For a detailed discussion of the role that counterfactuals and modal properties place in Fischer and Ravizza's theory, see McKenna (2013, 154).

pernicious, implicit biases), or she must explain how humans can still be reasons-responsive, even when they think and act from nonconscious, intuitive processes. The remainder of this essay is my attempt to effect this latter option.

3. Dual process Theories and Haidt's Social-Intuitionism

According to dual process theories, the human mind operates by way of two distinct types of processes. On one standard description, "type 1" processes are fast, automatic, associative, nonconscious, and affective, and "type 2" processes are slow, controlled, rule-based, conscious, and cognitive (Kahneman 2003; Frankish 2010). This picture has been challenged in recent years—something I'll discuss in Section 4—but it serves well enough as an initial description.

The idea that the human mind is "partitioned" can be traced back as far as Plato. In Plato's evocative chariot allegory, the rational part of the soul literally reins in the spirited and appetitive parts. And it is quite common, at least in the western philosophical tradition, to value the rational part of the human mind over its (unreliable) passionate and instinctual aspects, and to attribute most of humanity's ills to our frequent failure to use the former to control the latter. That we should exercise this sort of rational control over our thought and behavior has, with a few notable exceptions, typically been revered as something of an "Ur-responsibility."

What is potentially threatening about recent dual process theories of the human mind, then, is not this partitioning as such, or their claim that certain parts of the mind are less reliable than others, but is rather the doubt they cast on the *efficacy* and *scope* of rational control. So, while a dual process theory is simply any empirical theory about human psychology which posits two such distinct types of processes, the challenging findings of the last 50 years or so have been *how little control* the latter have over the former, and *how much* of human behavior takes place outside of the scope of that control. As Jonathan Haidt summarizes in his landmark essay, "The Emotional Dog and It Rational Tail" (2001): "The affective system has primacy in every sense: It came first in phylogeny, it emerges first in ontogeny, it is triggered more quickly in real-time judgments, and it is *more powerful and irrevocable when the two systems yield conflicting judgments*" (Haidt 2001, 819; emphasis mine). In a previous study, Haidt and colleagues (Haidt, Björklund & Murphy 2000) found that participants were likely to judge certain "harmless" taboo violations to be wrong—from incest to masturbating with a chicken carcass—despite being unable to *justify* (to *give reasons for*) that judgment, a phenomenon they dubbed "moral dumbfounding." Haidt takes these and other similar findings to support the conclusion that the vast majority of our moral judgments are caused by intuition, and that when we do engage in explicit deliberation about morality, it primarily serves to provide post hoc rationalization of these intuitive judgments (confabulation).

On the social-intuitionist model Haidt defends, an individual's reasoning and private reflection does sometimes influence her judgment, but it is supposed to be rather unusual.

On the other hand, Haidt is confident that reasoning can play a significant causal role in moral judgment when it "runs through other people", which he calls the "reasoned persuasion link" (Haidt 2001, 819). Arguably, this is something of a misnomer, since Haidt goes on to clarify that reasoned persuasion "works not by providing logically compelling arguments but by triggering new affectively valenced intuitions in the listener" (Haidt 2001, 819). Similarly, the "social persuasion link" is said to play a significant role in determining an agent's moral judgments, but again, not because of *reasons.* Rather, it is the agent's attunement "to the emergence of group norms"—which Haidt glosses as the agent's conformity to her friends', allies', and acquaintances' moral judgments (Haidt 2001, 819). This is why Haidt's is a *social-intuitionist* model: moral judgments are mostly caused by an individual's intuitions, including those she comes to have through social interaction.

Can human beings be *blameworthy* for failing to recognize and respond to reasons, particularly when they act, form attitudes, or make judgments as a result of intuition? The most pressing cases are surely those which pertain to agents' implicit biases concerning race, gender, sexuality, and, in general, biases which involve "negative evaluative tendencies directed towards people based on their membership in a stigmatized social group" (Washington & Kelly 2016, 17). But the issue is potentially thoroughgoing: if human agency operates largely by way of nonconscious, intuitive processes, then we may not be responsible for much of anything we do. In what follows I will work within the assumption that Haidt's and other dual process theorists' conclusion about the causal priority of nonconscious, affective, and intuitive processes is *correct.* This means I am going to assume that most human judgment is the result of such processes, and that private, explicit deliberation primarily serves to rationalize these intuitive judgments.[6] My aim will be to show what this conclusion, if true, does and does not say about the viability of just one, albeit highly influential approach to responsibility.

4. Intuitions as Reasons-Responses

As I discussed in Section 2, Fischer and Ravizza most commonly characterize the capacity to recognize and respond to reasons using a perceive-think-act model of human agency, in which actions are mediated by conscious, practical reasoning. If the dual process model is correct, then this is only rarely what happens. The most obvious first move for the reasons-responsiveness theorist, then, would be to argue that recognition and response to reasons need not be mediated by conscious deliberation at all. The capacity to recognize and respond to reasons would instead be understood along what might be called a perceive-*process*-act model of agency, where the processing in question need not take place within, or even be

[6] Of course, it is debatable whether intuitive processes really have this kind of priority and predominance in judgment. For example, Steven Clarke (2008) argues that the empirical evidence offered in favor of this claim is not decisive. However, my aim in the present paper is to argue *from* the truth of this priority claim to the conclusion that agents may still be justifiably held accountable on a reasons-responsiveness approach.

accessible to, conscious awareness. Supporting this idea, Fischer and Ravizza argue that their account "applies naturally and smoothly to nonreflective mechanisms of various kinds", once we dispense with thinking that recognizing and responding to reasons *requires* conscious deliberation: "The mere *recognition* that certain reasons exist does not imply that the agent is considering and weighing those reasons as part of an attempt to answer the practical question at hand" (1998, 87).

Reasons-responsiveness could thus be understood in a more expansive sense: as a capacity to perceive practically or normatively salient features of a situation and be suitably motivated by those perceptions—i.e., to respond in the ways those features prescribe and avoid responding in ways they proscribe. Importantly, these suitable motivations may take the form of strong, intuitive "gut feelings," the reasons for which may not be introspectively accessible to the agents who experience them, *even though* the agent really is recognizing and responding to some such reasons.

After all, the causal priority which dual process theorists attribute to affective processes does not, on its own, say anything about whether these processes can be reasons-responsive in this more expansive sense. However, they do tend to be described as biased and unreliable (Kahneman 2003, 2011; Greene 2007, 2013). So, even if we allow that such processes sometimes pick up on practically and normatively salient features and guide agents to act in ways that those features require or support, they may do so only rarely and be subject to falter in a wide range of scenarios.

Contrary to this trend, Peter Railton has argued that affective processes can be reliably responsive to reasons, due to the manner in which these processes *learn* and the kind of *cognitive resources* they exploit. In his view, some of these processes are "smarter," and the intuitive judgments produced by them are more reliable, than standard dual process models have suggested. He has, for instance, argued that the intuitive responses of participants in studies like Haidt's may reflect robust causal information about the world and the likelihood of certain harmful consequences. A particular instance of risky behavior, such as incest or playing Russian roulette, might not have had any negative consequences, but the gut feeling that such behavior is to be avoided in general arises from the fact that usually there's a good chance that it would (Railton 2014; Stanley, Yin & Sinnott-Armstrong 2019).

Railton's argument builds on recent developments in affective and computational neuroscience, which suggest that some of our intuitions result from processes that detect and encode statistical information about the environment in the form of causal models or maps—rather than operating exclusively by way of automatic, inflexible, and associative heuristics. This has led to a revised distinction between type 1 and type 2 processes in terms of the *learning mechanisms* by which they operate: "model-free" learning in the case of type 1 processes, and "model-based" learning in the case of type 2 processes (Crockett 2013; Cushman 2013). I find Joshua Greene's summarization of this distinction particularly helpful:

> Model-based learning involves accumulating information about the decision environment and using that information to build a causal model of that environment. For example, a rat in a maze might learn to obtain a reward by exploring the maze and building an internal map of the maze, which includes the location of the reward. […] Model-based learning and decision-making corresponds to what we would naturally identify as reasoning and planning: using an understanding of how the world works to identify a sequence of actions that will get one to one's goal.
>
> Model-free learning and decision-making work in a fundamentally different way. Instead of building an explicit model of the world, model-free learners attach positive or negative values directly to actions (or action-context pairs) based on whether and to what extent those actions have been rewarded in the past. For example, if a rat stumbles upon the rewarding cheese after making a right turn out of a red room, the next time it finds itself in the red room (or a similar room) it will feel an urge to turn right. (Greene 2017, 69)

Now, while model-based decision-making is characterized as corresponding "with what we ordinarily recognize as reasoning" (Cushman 2013, 277), we should be careful not to assimilate it with conscious deliberation. Rather, model-based learning mechanisms are said to correspond with ordinary reasoning because of the way they process information: i.e., in consultation with internal representations of the broader environment and the consequences of interacting with that environment, which enable the organism to engage in complex and projective means-end reasoning. But importantly, the organism need not consciously represent its causal map of the world, or even be able to do so, in order to reason about how best to explore that world in order to achieve the things it values (Railton 2017, 176). Railton takes this to support a more optimistic picture of intuitive judgments, since it suggests that, even where we may be unaware of the reasons behind our judgments, we may nonetheless be recognizing and responding to such reasons.

But in what way, precisely, does this distinction support the claim that humans may be responding to reasons even when they act from nonconscious, intuitive processes? Is the idea that processes which utilize causal maps (type 2 processes) are reasons-responsive, whereas those which rely on associative expectation values (type 1 processes) are not? Not necessarily. If we embrace the more expansive notion of reasons-responsiveness glossed at the start of this section, then either type of process *can* be responsive to reasons. To draw on the passage from Greene (2017) above: In a world in which all red rooms have cheese to the right, a rat which associates intrinsic value with turning right in red rooms will get along just fine. Indeed, in such a world, and for such a creature, *being in a red room* constitutes a reason to turn to the right.

What is really at issue in Railton's discussion is the *reliability* of type 1 and type 2 processes, given that the actual value of certain responses is often subject to change. An essential difference between model-free and model-based learning mechanisms is the relationship

each bears to *corrective feedback,* or, differently stated, to the *selection pressures* which shape these mechanisms over time—and, consequently, the thoughts, feelings, and behaviors they produce. It is not as though model-free learning mechanisms are completely static stimulus-response relationships. They, too, are constantly updating to reflect the reward values of specific responses as those values change over time. But this updating procedure takes place slowly and is susceptible to certain errors. For example, in "devaluation procedures" (Greene 2017, 70), a rat will continue to respond to its situation in ways which have been associated with high expectation values (like pressing a lever which releases a food pellet), even when the rat no longer has the relevant desire (is no longer hungry), or even when the actual value of that response has changed (e.g., the food has been poisoned).

Model-based learning mechanisms, by contrast, display a measure of diachronic flexibility which is simply not available to their model-free counterparts. This is because the values represented in models are connected not with <action, situation> pairs, but rather with the *consequences* of such pairs. "A model-based algorithm, in contrast, has the capacity to recognize that the specific outcome associated with pressing the lever is food," making it possible for the rat to update the value of pressing the lever to reflect its satiated state (Cushman 2013, 279). Model-based learning is thus characterized by an *in-order-to* structure that reflects causal relationships in the world, whereas model-free learning contains this causal information only *implicitly* in the intrinsic values associated with specific actions. So, while the responsiveness of either type of process is largely a function of the relationship between (a) the environment in which it is presently operating and (b) the environments which shaped them, model-based processes are *more likely* to be reasons-responsive in a changing world because they can flexibly reduce the discrepancy between these two environments by updating the models.

However, a central problem remains. Despite agreeing with the broader psychological picture that Railton endorses, Greene has argued that Railton's optimistic view of moral intuition fails to address what he calls the problems of "bad training" and "bad data" (Greene 2017, 72-5). In Greene's view, even those intuitions which are caused by sophisticated, model-based learning mechanisms are liable to mislead us if the experiential samples from which their models have been drawn are themselves biased, which they often are. Consider, as just one example, the kind of causal models an individual is likely to have if her primary exposure to members of other racial and ethnic groups has been mediated by news sources which represent members of those groups almost exclusively in connection with violent crimes.

This is also why the reasons-responsiveness approach cannot simply rely on the claim that nonconscious, intuitive processes have the *potential* to be reasons-responsive. For even if this is true, the really important question is whether, when these processes *fail*, individuals are responsible (can be *blamed*) for those failures. If the reliability even of model-based intuitive processes is contingent on fortuitous learning environments, then it is at least not obvious why these failures should *count against the agent* in the way that is required for

blameworthiness. Blame is a charge to the effect that someone could have and should have done something, but didn't *and has no excuse*. The remainder of this essay will explain how the reasons-responsiveness approach can handle this problem.

5. Reasons-Responsiveness and Ecological Control

Human beings bear a unique relationship to their environments. Individuals are born into a world already replete with *cumulative culture*—a vast repository of intellectual and technological resources built up by their predecessors (Richerson & Boyd 2005). One form these resources can take are empirical studies about human psychology, as well as effective strategies for mitigating things like implicit biases. Moreover, some philosophers have argued that the mind "extends" into the environment, recruiting stable features of the environment to "off-load" certain cognitive processes (Clark 2007). Some of the more promising strategies for mitigating limiting aspects of human psychology, such as implicit bias, may involve shaping the human environment in ways that beneficially shape us, in turn.

An especially pressing concern raised by the broader dual process model is that human beings are often unaware of their implicit biases, and even once they become aware of them, they may not be able to directly control their effects. This suggests, for example, that many racist beliefs are held unknowingly and unintentionally, and that, despite an individual's explicit rejection of these attitudes and the behaviors they guide, she may not be able to help the fact that she has and is guided by them. Responding to this concern, Natalia Washington and Daniel Kelly (2016) argue that agents can still be blameworthy for their implicit biases "when knowledge about such mental states [and about how to regulate their effects] is available in her epistemic environment" (Washington & Kelly 2016, 13). Similarly, Jules Holroyd and Daniel Kelly (2016) argue that actions which result from implicit biases can be attributed to agents in the way required for moral evaluation, and perhaps even for blame, so long as the agent could have exercised "ecological control" (Clark 2007) over those biases and their effects.

Both articles thus urge theorists (and practitioners) of responsibility to place less importance on introspectively available knowledge and direct, conscious control, and to place more importance on the epistemic and regulative resources available in the individual's environment. The solitary individual may not have what it takes to regulate her implicit biases, but the individual-plus-environment does. "Today, the amount of empirical evidence collected on implicit biases is enormous, and it continues to mount. Much more is known in general, and that knowledge is much more widespread in [today's] environment than it was in the early 1980s" (Washington & Kelly 2016, 24). Thus they claim that, because of this difference in *external context*, someone alive today should already be aware of implicit bias in general, and of her own implicit biases in particular, whereas the same cannot be said of someone living in the 80s. Washington and Kelly apply a similar line of reasoning to control-related excuses:

> For not only does an individual need to know that she has implicit biases before she can even try to exert control over them, but doing so consistently and effectively will also require a special kind of knowledge—specifically, knowledge of and facility with the kind of techniques and methods that are being shown to be effective by the empirical research on the malleability of implicit bias. (Washington & Kelly 2016, 25-26)

An agent's inability to directly control her implicit biases is not an excuse, then, so long as she could have already learned about and practiced techniques for correcting those biases—or, at least, prevent them from influencing her behavior. For example, the members of a hiring committee could have removed the names of job applicants from their résumés beforehand to prevent themselves from favoring applicants with "white-sounding" names.

Similarly, Jules Holroyd and Kelly (2016) argue ecological control can help explain why agents can be morally evaluated, and perhaps even blamed, for their implicit biases. "A person might engineer her 'external' epistemic environment in other ways to ensure that her intentions and values are more fluidly expressed in her actions and judgements, and not distorted by the operation of implicit biases" (Holroyd & Kelly, 121-22). One of the empirically supported examples they mention is surrounding oneself with counter-stereotypical images, such as images of admired black celebrities. The use of such "environmental props," and in general the availability of information about effective strategies for mitigating implicit bias, is taken to support their conclusion that "the idea that an agent's implicit biases are beyond her control in any relevant sense is simply false" (Holroyd & Kelly, 123).

6. The "Hard Problem" of Responsibility

The idea that ignorance and lack of control do not always excuse moral failings, particularly when the agent's ignorance or lack of control can be traced back to factors which the agent is responsible for, is a familiar one in the responsibility literature. What is novel about Washington, Holroyd, and Kelly's discussions is the capacitating role they attribute to an agent's environment. This is what supports their claim that agents do not need to be introspectively aware of their biases and do not need to have the capacity to exert direct control over those biases in order to be responsible for them. That is, an agent who fails to respond to reasons today, because she acts from unreliable, biased processes, is still blameworthy for that failure, so long as her environment was such that she could have taken steps to learn about and regulate those processes in the past, but nonetheless failed to do so. They could have, and should have, known better.

But I am doubtful that Greene's skepticism is adequately addressed by this appeal to the agent's prior failures. Our contemporary environments do contain information about these processes and strategies for regulating them, but this alone may not be sufficient to *fault* an agent for failing to seize these learning opportunities. In keeping with the reasons-responsiveness framework, for someone to be responsible for this kind of failure, she must

have had the capacity to recognize and respond to the reasons there were, on some previous occasion, to learn about and deploy these strategies. But if she did have this capacity, then what explained her failure to exercise it? There are two options: Either the factors which explained this prior failure are not, themselves, factors which she is responsible for, in which case they would count as *excuses;* or the agent *is* responsible for these factors, in which case we must again ask, what explains her failure to overcome them?

To make things more concrete, recall the example from Section 2 about Mariem, the driver who failed to yield to the pedestrians. Surely there is something that explains why Mariem failed to yield, and if she is blameworthy for this failure, then the factor which explains her failure must be something for which Mariem is responsible. We wouldn't say, for instance, that Mariem is responsible for her failure if it was caused by a sudden heart attack, or was due to mere chance—say, a neural misfiring or some other glitch that prevented the exercise of her normal capacities. Here's one possibility: Mariem just didn't feel like going through the motions on this particular occasion. Mariem, we are supposing, is weak-willed. She sometimes acts against her better judgment simply because she lacks the motivation to carry through with it. However, as McGeer and Pettit explain,

> (…) these explanations are special. They allow us to condemn the failure that they explain only because we hold the agent responsible for the persistence of the trait in question; that trait is not, as we might put it, a brute factor. We have to think, in accordance with the reason-responsive approach, that the agent has the specific capacity to respond to reasons and overcome that trait. We must deny, for example, that the laziness or weakness of will is sourced in some pathology, or even some pattern in the past, that makes it impossible to overcome without serious therapy or biochemical intervention. If we thought that the trait was maintained in that way, we would treat it as an excusing factor. (McGeer & Pettit 2015, 164)

The issue with explanations which appeal to character traits, then, is that they only push the explanandum back. For we must then ask, "But what explains the agent's failure to overcome *that trait?"* Suppose Mariem is weak-willed, and that, on some prior occasion, she had the capacity to recognize and respond to the reasons there were in that situation to overcome this trait. Then there must be something which explains her failure to exercise this further capacity and overcome the trait, and we must yet again consider whether it is a factor which she is responsible for. We can keep on in this vein, but we will either end up with a vicious regress of failure explanations, or we will end up with something which is *brute* in McGeer and Pettit's sense—that is, something which the agent lacked the capacity to overcome at the time (see Figure 1 below).

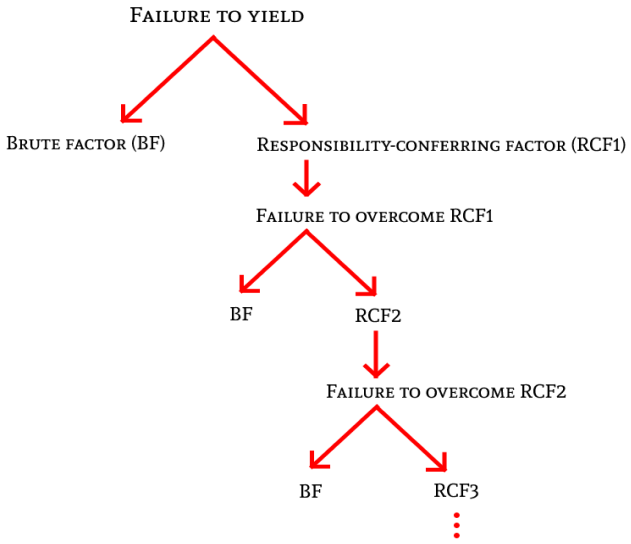

Figure 1. Explanatory regress on the traditional account.

The same general problem will emerge when explaining an agent's responsibility for her implicit biases. Suppose Judie hired a candidate with a white-sounding name, rather than a better-qualified black candidate, because of her implicit racial biases. When she made this decision at time, *t*, she was not responsive to reasons (e.g., the reasons there were to hire the best qualified candidate), because of her unchecked implicit bias. In Washington, Holroyd, and Kelly's view, Judie is still blameworthy for this failure, because she could have (and should have) known that she has this implicit bias, and that an effective way to prevent it from affecting her hiring decision would be to remove the names of the applicants from their résumés beforehand. Seems right to me. But then it should be true that, at some previous time, *t-1*, Judie (a) had good reasons to learn about and deploy strategies for regulating her implicit bias, (b) had the capacity to recognize and respond to those reasons, but (c) failed to exercise that capacity.

Suppose Judie met conditions (a)-(c) at *t-1*.[7] Now we have to contend with one further possibility: that Judie's failure to exercise her capacity at *t-1* is explained by an *excusing factor.* Surely something explains her failure—she didn't "just" fail to make use of these resources—and the factor which explains her failure cannot be something which Judie had no control over, or else Judie won't be blameworthy for her biased hiring decision at *t.* Suppose it is not an excusing factor. That is, suppose Judie at *t-1* met conditions (a)-(c) with respect to this factor. She had the capacity to recognize and respond to the reasons there were to overcome

[7] If Judie at *t-1* did not have the capacity to recognize and respond to those reasons, she could still be blameworthy, but only if there was some previous time, *t-2*, at which Judie met conditions (a)-(c). This is just the same tracing procedure as before: it grounds Judie's blameworthiness at *t* in her failure at *t-2* to take the necessary steps to regulate her bias.

it, but she failed. But this just starts things all over. For surely something explains this distal failure. Now we'll need to see whether the factor which explains the distal failure is an excusing factor. It seems that either we end up with an infinite regress, or Judie's biased hiring decision at *t* is ultimately due to an excusing factor.

Surely this is a worry only a philosopher could have. No one seriously doubts, *in situ,* that Judie couldn't have known to remove the names from the applications. If Judie tried to excuse her hiring decision in this manner—just imagine!—she'd be met with very little patience. But if the reasons-responsiveness approach is to explain Judie's blameworthiness, it has to solve this problem.

McGeer and Pettit's solution to this problem begins with what they call the "developmental assumption": that we largely owe it to others—to our past and ongoing interactions with other people in our moral community—that we are responsive to moral considerations. And from this assumption they aver that many of us probably come to be intrinsically motivated to care about how other people (particularly those whose moral authority we recognize) feel about us and about the things we do.[8] So far this may not seem helpful—certainly Judie's responsiveness moral considerations, imperfect as it is, is contingent on previous interactions she had with whomever raised and educated her. That is—*of course* Judie's standing disposition to respond to those reasons has a social-historical origin. The important upshot of the developmental assumption, however, is not what claims about Judie's past, but is rather what it claims about her current and ongoing sensitivity to feedback from other people.

McGeer and Pettit propose that we think of an agent's capacity to recognize and respond to reasons as a product of two sensitivities: the agent's *standing sensitivity* to the reasons (think of this as the likelihood that she will respond to reasons 'on her own'), and her *situational sensitivity* to others' expectations. This latter sensitivity may fruitfully be characterized as a second-order sensitivity in that it functions to modulate the first, usually strengthening it:

> Suppose you bring to a choice a sensitivity to reasons of a certain strength, S, where the strength of a disposition is determined by the probability it puts in place that under a relevant scenario or stimulus you will respond to reasons. The idea is that your sensitivity to audience in that choice may reinforce your sensitivity to reasons by making you more attentive, more careful, more motivated to track the reasons that there are, at least for the duration of the choice. It may increase the strength of that disposition so that your ultimate responsiveness to reasons is of strength, S-plus, not just S. (McGeer & Pettit 2015, 172)

[8] To say we are intrinsically motivated in this way is just to deny that we care about others' expectations of us for merely instrumental reasons, such as the inconveniences and prudential set-backs we would face if we lost their respect or, indeed, their concern for us altogether. Rather, in seeing others as our authorized moral audience, we experience their expectations of us as salient in their own right.

But even if we accept the revised, two-tired capacity to respond to reasons, isn't it still true that, when Judie failed to recognize and respond to reasons on this particular occasion, it was ultimately due to some brute factor or other? Actually, yes. But recall that the problem with this brute, failure-explaining factor was that we lacked an explanation for why it should not count as an excuse—why we are still justified in blaming Judie, even if her failure is traceable to such factors. The payoff of the proposed revision is that it specifies the conditions under which such brute factors are *excusing*:

> (…) excuses are just those failure-explaining factors of which the following is true: according to assumptions encoded in our injunctive practice—these may vary, of course, across cultures—there is little hope of neutralizing their effect by holding people responsible in their presence. And so, on that theory, the features that explain failure without counting as excuses are just those factors—those glitches and chances—that are susceptible, according to our injunctive assumptions, to the regulatory effects of our holding one another responsible. (McGeer & Pettit 2015, 183-84)

Judie is eligible for blame, then, precisely because blaming her helps her to regulate that brute, failure-explaining factor. McGeer and Pettit's suggestion, then, is that we take a "forward-looking" approach to responsibility, where practices like blame are partly justified by the effects they are likely to have on the agent's future thought and behavior.

In essence, the objection that I am using the "Hard Problem" to leverage against Washington, Holroyd, and Kelly is that they don't go far enough. Specifically, although they appeal to the capacitating role played by an agent's environment, they don't consider the place that social practices like praise and blame have in that environment. An agent's capacity to recognize and respond to reasons is partly enabled *by us*—by the rest of us, who stand in relations of influence to that agent. When a community blames someone like Judie for her biased hiring decision, that response is itself part of the environmental scaffolding in virtue of which Judie is responsive to reasons.

Recall that the problems of "bad data" and "bad training" is that agents' intuitive processes very often *are not* reasons-responsive, and that their lack of reasons-responsiveness is due to contingent learning histories. According to McGeer and Pettit, an agent's sensitivity to others functions as a second-order sensitivity, in that it augments the strength of the agent's first-order responsiveness to reasons. This allows us to explain why agents may be held morally responsible for some (though surely not all) brute failure-explaining factors.

> If your responsiveness to reasons in a given choice is a function of two forces, then naturally it becomes possible for your responsiveness to result from different combinations of those forces. The two sensitivities may combine in different measures to produce responsiveness and any degree of responsiveness may be realized via any of a range of equivalent combinations. […] [W]hat we must now notice is that when I take you to be responsive, it may be that I do not credit you with a very reliable, standing

> capacity to respond to reasons. I may take you to be suitably responsive—to have the required capacity—only in the actual or foreseen presence of the audience that I and perhaps others constitute. (McGeer & Pettit 2015, 172, 173; emphasis mine)

Agents are fit to be held responsible for their moral failings because they are capable of responding to reasons, where they have this capacity partly because they are involved in a wider system of social practices. Their standing sensitivity may even be relatively low, but if in combination with the sensitizing effects of a moral audience they are *rendered capable* of acting as the reasons require, then it is appropriate to hold them responsible.

7. A Way Forward for Reasons-Responsiveness

On the resulting picture, we are able to maintain the central claim of the reasons-responsiveness theory—that an agent is responsible for something she does only if she acts from her own reasons-responsive process—but resist the troubling inference that, if something like Haidt's social intuitionism is correct, this would include only a very narrow range of human thought and behavior. Broadly put, an agent is still responsible for what she does as a result of nonconscious, intuitive processes if they are processes which can be attuned to reasons (or whose effects can be regulated) through our accountability practices.

What is novel about this answer to the challenge posed by some dual process theories is that it identifies an important relationship between the "trainability" of processes like implicit biases—that these processes or at least their effects can be regulated so as to strengthen responsiveness to reasons—with the capacitating role that our accountability practices themselves play in that process of training. As I argued in Section 3, it is not enough to show that intuitive and affective processes sometimes reliably track practically and normative significant features of the world and guide thought and behavior in ways that those features make appropriate or require. Greene rightly indicates that these processes are, in a sense, at the mercy of their learning histories, which may often be tainted by "bad data" and "bad training". In Section 4 I considered one way of abetting this problem—by appealing to ecological factors which the agent might avail herself to, both to learn about, and to learn strategies for regulating, processes like those that result in implicit racial biases. In other words, agents can and should avail themselves to better data and better training. But as I argue in Section 5, accounting for an agent's blameworthiness in this way only passes the buck on to failures the agent made in the past to discover and learn from these resources. To justifiably blame the agent for the consequences that these failures have had on her subsequent thought and behavior—neglecting to omit the names on resumes for a hiring search, for example—the failures will, in turn, need to be explained in a manner that confers responsibility onto the agent. I argue that this can be done if we recognize that our social practices of accountability are themselves a capacitating feature of an agent's ecology, in the manner suggested by McGeer and Pettit's solution to the "Hard Problem". If we recognize that the degree to which an agent is responsive to reasons is partly dependent on that agent's

involvement in a system of accountability practices, then whether it is justifiable to hold an agent accountable is partly a function of the effect this sort of practice would have on the agent.

There are two important clarifications to be made about the resulting view. The first is that, while the effects of our practices play a justificatory role, they do not justify those practices by appeal to *desirable consequences.* It is not merely that expressions of praise and blame will bring about valuable results, such as encouraging better behavior. Rather, it is that expressions of praise and blame develop the very capacities which agents need in order to be justifiably held to certain expectations or norms. An agent's responsibility depends on her capacity to reliably recognize and respond to reasons, where the strength of this capacity rises or falls depending on whether the agent participates in a community with these sorts of practices. In this way the account is not strictly a consequentialist account. Reasons-responsiveness is not merely good we have reason to promote; it is also what makes humans capable of, and thereby subject to, legitimate norms and expectations.

The second clarification concerns the specific nature of the effects. It is useful here to contrast an alternative response to Haidt's social intuitionism. Steven Clarke (2008) argues that there are a number of ways people might be influenced to rely more heavily on reasoned, conscious deliberation, rather than intuition, and discouraged from falling into post hoc reasoning. As examples he mentions alternative schooling styles, deferring our opinions to the judgment of experts, critical thinking skills, and social and intuitional changes, such as adjusting the rules of public debates. On the one hand, many of these suggestions are in line with the view I have presented, inasmuch as they utilize wider social circumstances to augment the agent's capacity to recognize and respond to reasons. However, Clarke's suggestion is to encourage the agent to *switch* to reasoned deliberation. I am not opposed to this sort of strategy, and I agree with Clarke that more empirical evidence is needed before we can be sure that strategies like these wouldn't be effective enough to undermine Haidt's claim about the priority of intuition. But my aim has been to work within the constraints of that claim and to show how a reasons-responsiveness theory of responsibility could be preserved even if it turns out to be decisively true. So, the capacitating effect I have in mind is not a switch to conscious deliberation; rather, it is the regulation of the agent's intuitive responses themselves, so that these might be made more reliable. Also, institutionally-geared strategies similar to those Clarke mentions may also be effective in this respect, but the view I have defended is concerned only with the effects of interpersonal practices, such as directed expressions of praise or blame.

8. Conclusion

To bring things to a close, I want to direct our attention back to Haidt's social-intuitionist model. According to Haidt, even though reasons play a surprisingly insignificant role in the mental lives and behavior of human beings, social interactions can have substantial effects

on the intuitions which guide our thought and behavior. The challenge which dual process models present for the reasons-responsiveness approach was that they cast doubt on the scope and efficacy of an individual's private, conscious, rational control. If the capacity to respond to reasons is supposed to be mediated by that form of control, then the truth of a social-intuitionist model like Haidt's would be very difficult to square with the reasons-responsiveness approach. But there's good reason to think that the capacity to recognize and respond to reasons can be realized in our nonconscious, intuitive processes, and that the reliability of these processes can be "trained up" by social interactions. Of course, they can also be badly trained by those interactions and by unrepresentative learning experiences, but that is *all the more reason* to engage one another, holding each other to higher standards than we would otherwise be able to meet. This is why I think notions like ecological control are so promising, but can also be misleading if we forget that a central component of human ecology, indeed, a piece of social technology passed down through cumulative culture, are our responsibility practices. The justifiability of those practices is to be found, in part, in the function they serve to train the very capacities on which their justifiability depends.

References

Clark, A. (2007). Soft selves and ecological control. In D. Spurrett, D. Ross, H. Kincaid, L. Stephens (Eds.), *Distributed Cognition and the Will*, pp. 101-122. Cambridge: MIT Press.

Clarke, S. (2008). SIM and the City: Rationalism in Psychology and Philosophy and Haidt's Account of Moral Judgment. *Philosophical Psychology, 21*(6), 799-820. htt ps://10.1080/09515080802513250

Crockett, M. J. (2013). Models of Morality. *Trends in Cognitive Sciences*, *17*(8), 363-366. https://10.1016/j.tics.2013.06.005

Cushman, F. (2013). Action, outcome, and value a dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273-292. https://10.1177/1088868313495594

Fischer, J. M., Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility.* Cambridge: Cambridge University Press.

Frankfurt, H. (1969). Alternative Possibilities and Moral Responsibility. *Journal of Philosophy, 66*(23), 829-839. https://10.2307/2023833

Frankish, K. (2010). Dual process and Dual-System Theories of Reasoning. *Philosophy Compass, 5*(10), 914-926. https://10.1111/j.1747-9991.2010.00330.x

Greene, J. (2017). The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition* 167, 66-77. https://10.1016/j.cognition.2017.03.004

Greene, J. D. (2007). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Volume 3: The Neuroscience of Morality: Emotion, Disease, and Development,* pp. 35-79. Cambridge: MIT Press.

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review, 108*(4), 814-834. https://10.1037//0033-295X.108.4.814

Holroyd, J., Kelly, D. (2016). Implicit Bias, Character, and Control. In A. Masala, J. Webber (eds.), *From Personaliy to Virtue: Essays on the Philosophy of Character,* pp. 106-133. Oxford: Oxford University Press.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697. https://10.1037/0003-066X.58.9.697

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Macmillan.

McGeer, V., Pettit, P. (2015). The Hard Problem of Responsibility. In D. Shoemaker (ed.), *Oxford Studies in Agency and Responsibility, Volume 3,* pp. 160-188. Oxford: Oxford University Press.

McKenna, M. (2013). Reasons-Responsiveness, Agents, and Mechanisms. In D. Shoemaker (ed.), *Oxford Studies in Agency and Responsibility, Volume 1,* pp. 151-183. Oxford: Oxford University Press.

Railton, P. (2017). Moral Learning: Conceptual foundations and normative relevance. *Cognition, 167*, 172-190. https://10.1016/j.cognition.2016.08.015

Railton, P. (2014). The affective dog and its rational tale: Intuition and attunement. *Ethics*, *124*(4), 813–859.

Richerson, P., y Boyd, R. (2005). *Not by Genes Alone: How Culture Transformed Human Evolution.* Chicago: The University of Chicago Press.

Smith, M. (2003). Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion. In S. Stroud, C. Tappolet (Eds.), *Weakness of Will and Practical Irrationality*, pp. 17–38. Oxford: Clarendon Press.

Stanley, M. L., Yin, S., Sinnott-Armstrong, W. (2019). A reason-based explanation for moral dumbfounding. *Judgment and Decision Making, 14*(2), 120-129. URL: https://search.proquest.com/docview/2200763447?accountid=13360

Vargas, M. (2013). *Building Better Beings: A Theory of Moral Responsibility.* Oxford: Oxford University Press.

Washington, N., Kelly, D. (2016). Who's Responsible for This? In M. Brownstein, J. Saul (eds.) *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics,* pp. 11-36. Oxford: Oxford University Press.

Wason, P. C., J. St. B. T. Evans (1975). Dual Processes in Reasoning? *Cognition*, *3*(2), 141-54. https://10.1016/0010-0277(74)90017-1