LEARNING RESPONSIBLY:
ESSAYS ON RESPONSIBILITY, NORM PSYCHOLOGY, AND PERSONHOOD

by

Stephen A. Setman

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Philosophy

West Lafayette, Indiana

August 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Daniel Kelly, Chair

Department of Philosophy

Dr. Michael Jacovides

Department of Philosophy

Dr. Taylor Davis

Department of Philosophy

Dr. Victoria McGeer

Department of Philosophy

Princeton University / Australian National University

Approved by:

Dr. Christopher Yeomans

Dedicated to Dillon James McGinn.

ACKNOWLEDGMENTS

I am grateful and frankly lucky to have known so many excellent and caring people in my life, academically as well as personally. Of some I can sincerely say that, if I hadn't known them, I would be a very different person living in very different circumstances. Mark Graham, who was my high school teacher for four years at Brandywine, and also one of the most passionate and capable teachers I have ever had, taught me how to write, how to think critically and rigorously about my life and my world, and why it was important to do so. Around the same time my dear friend, Phil Calafati, guided me through the college admissions process and alerted me to some of the reality of that process and what it would take to get through it with success. At my undergraduate alma mater, Gettysburg College, I quickly found my home in the Philosophy Department. Steve Gimbel, Kerry Walters, Dan DeNicola, Lisa Portmess, and of course my senior thesis adviser, Gary Mullen, taught me philosophy and modeled for me how to teach philosophy. In different ways each of them continues to be an example of teaching excellence that I regularly draw inspiration from in my classroom. Special thanks to Gary for taking me on as an undergraduate research assistant, and for advising me to think of the value of graduate school independently of its professional upshot. Vern Cisney, who also earned his Ph.D. at Purdue and who is now Associate Professor in Interdisciplinary Studies at Gettysburg, was the person who recommended I apply to Purdue's philosophy program. Kerry Wallach, Associate Professor in German Studies, was there for me at some emotionally difficult times, advised me while I applied to graduate schools, and has been an honest and encouraging mentor on many occasions since. At Purdue there are many I have to thank for the education I received in the theory, practice, and instruction of philosophy. I would especially like to express my gratitude toward Paul Draper, Bill McBride, Jackie Mariña, Lynn Parish, Chris Yeomans, and the members of my dissertation committee, Taylor Davis, Mike Jacovides, and my dissertation chair, Dan Kelly. If there's something that being a student for 22 years has taught me, it is that learning who to learn from is absolutely vital. I feel very fortunate to have found the particular sort of counsel I needed in the members of my dissertation committee: Taylor's healthy skepticism and no-nonsense empiricism; Mike's thorough, sobering, and on-point feedback; Dan's fun and ecumenical attitude, his sharp wit, and his indefatigable commitment to guiding his students and

giving them the opportunities they need to develop as researchers. I also want to thank Victoria McGeer, whose philosophical work has been so very influential, for agreeing to be the outside reader on my committee. Many thanks as well to Administrative Assistant, Vickie Sanders, for helping me navigate through all of the forms and deadlines, and for always being such a kind and welcoming presence in our department. On a personal note, I want to say how grateful I am for the incredible support network and community I have had in my colleagues, friends, and neighbors: Alzbeta, Luke, Dan, Richard, Diaz, Abby, Evan, Giuseppe, Gipsy, Scott, Kevin, Caleb, George, Charles, Kurt, Cameron, and so many others. I also want to thank my partner, Kariny. There is so much in you that I look up to, so much that impresses and inspires me, and so much that reminds me of what I have yet to learn and what I should be more grateful for. You have been the compassion and the love and the hope and the positivity that I needed, on so many occasions, to cut through the doubt and the denial. Thank you. Finally, I want to thank my parents—my dad, for his lightheartedness, amicability, and patience; my mother, for her generosity, and for giving me the freedom to discover myself.

TABLE OF CONTENTS

ABSTRACT

This dissertation argues for a number of theses related to responsibility, norm psychology, and personhood. Although most of the papers argue for "standalone" theses, in the sense that their truth does not depend the truth of the others, the five papers collectively illustrate a broader view of humans as (a) responsible agents who are (b) self-governing and (c) equipped with a capacity for norms, and whose agency (d) centers on dynamic responsiveness to corrective feedback. Drawing on this broader picture, the dissertation sheds light on ethical questions about our social practices and technologies, as well as descriptive questions about the nature of substance use disorder.

Most centrally, the dissertation argues that forward-looking considerations are relevant for responsibility, not merely because the consequences of our responsibility practices are desirable, but primarily because of a connection which I argue exists between relationships, norms, and learning. On the view I defend, an agent is a responsible agent only if she can learn from being held responsible, so as to regulate herself according to norms of which she presently falls short. I argue that, if it were not for the capacity of humans to learn from *social corrective feedback*, such as normative responses like praise and blame, humans would be unable to participate in norm-governed relationships and communities. It is in virtue of their participation in these relationships and communities that humans are subject to interpersonal norms, such that they can fulfill or violate these norms and be praiseworthy or blameworthy for doing so. So, without the kind of learning that makes participation in these relationships a possibility, humans could never be praiseworthy or blameworthy for anything that they do.

The dissertation also argues that human norm psychology has implications for how we should relate to "social robots"—artificial agents designed to participate in relationships with humans. I argue that, like humans, social robots should be equipped with a capacity to recognize and respond to normative feedback. Lastly, the dissertation resists a common narrative about addiction as being a form of akrasia in which agents act against their own better judgment. While this is certainly a central aspect of many cases of addiction, I argue that it fails to appreciate the ways in which addiction sometimes interacts with a person's identity and goals, especially in cases where the agent believes that the things she values would not be feasible if she did not continue to engage in addictive behavior.

INTRODUCTION

Human beings are responsible for what they do because they can learn—both in general, and specifically from being held responsible by others. This isn't the only reason humans are responsible for what they do, but it is a central reason, and one whose centrality has often been underappreciated or misunderstood. In this dissertation I explain why learning is at the heart of what justifies our practices of responsibility—paradigmatically, praise and blame—while at the same time preserving the intuition that when humans can be justifiably held responsible, it is because of what they have already done and how they have done it, and not merely because of desirable consequences. As I'll argue, if not for their capacity to learn from *social corrective feedback*, such as that communicated in normative responses like praise and blame, humans would be unable to participate in norm-governed relationships and communities. It is in virtue of their participation in these relationships and communities that humans can be legitimately subject to interpersonal expectations and norms, such that they can fulfill or violate these expectations and norms and be praiseworthy or blameworthy for doing so. Without the kind of learning that makes participation in these relationships a possibility, humans could never be praiseworthy or blameworthy for anything that they do.

When I say that humans are responsible because they can learn, the sense in which I mean it, or the type of responsibility I take myself to be referring to, may not be the only sense or type of responsibility imaginable. But it is certainly a type of responsibility which human beings have, given the sort of creatures they plausibly are. I believe it is also a type of responsibility which can serve as an adequate normative basis for justified practices of accountability—i.e., not just attributions of faulty character or faulty thinking, but also interpersonal expressions of blame and reparative demands. With respect to this question of justification, some responsibility theorists—and maybe some people in general—will never be satisfied with anything less than the most radical form of free agency imaginable, and the one for which there is also the least evidence. Perhaps some creatures out there possess this radical sort of freedom, and perhaps even humans possess it. But there is already so much richness and complexity in what we *know* about human agency, and it seems to strongly suggest that, in some relatively familiar sense of the term, humans are responsible for what they do.

Significance and Challenges

Merit, desert, accountability, reparations, and reform are all central features of human society. Philosophical work on responsibility thus has genuine practical significance. Even where we may not think of our practices in these terms, much of how we decide to relate to others, at the interpersonal, collective, and institutional levels, turns on considerations of responsibility. Whether and for what reasons a young adult is admitted to a university; why we blame others, from close friends to total strangers, when they fall short of important standards or norms; and, in general, our expectation that others take ownership of and correct harmful thought and behavior—all of these center on the basic question of what makes human beings responsible for what they do and have done.

As it turns out, answering that basic question is not completely straightforward. Consider the everyday sense in which most of us feel, intuitively, that what we make of our lives is genuinely up to us, at least within certain modest parameters. But upon closer inspection, many of us also recognize that the values and attitudes which frame our decisions are largely due to factors we did not choose. We, all of us, emerge out of historical conditions—familial, educational, economic, social, biological—for which we are not directly responsible. Moreover, as the cognitive and social sciences reveal more about human cognition and its sensitivity to environmental influence, we see that individual belief and motivation is not infrequently the product of processes and platforms over which the solitary individual has precious little control. Against these many "specters of determinism," how are we to maintain the intuitive sense of freedom and self-determination on which so many of our social practices depend for their legitimacy? Is merit recoverable from privilege? Desert from bad luck and bad influence?

In short: *Yes*. In this dissertation I argue that human responsibility and its associated practices—praise, blame, and the like—are not only recoverable, but are necessary tools for allaying those very concerns which have thrown human responsibility into question. According to the view I defend, these practices depend for their justifiability on an individual's possession of certain capacities of human agency, where these capacities depend for their development on the informational and motivational resources carried and communicated in those very same practices. Responsible agency is, in this sense, a type of skill. Like all skills, it must be developed and maintained through a process of (primarily social) learning, wherein feedback from the environment functions to cue agents into opportunities to hone their capacities, as well

as the normative standards and expectations around which they are to be honed. Just as a novice prep cook cultivates her culinary talents by studying alongside and under the caring—if also demanding—supervision of a practiced chef, so too must responsible agency be cultivated in human beings through example and instruction. On the view I defend, an agent is a *responsible agent* (is at least eligible to be held accountable for what she thinks and does) just in case she is capable of *learning* from those accountability practices—such as learning to see and be moved by considerations of normative significance, and learning how to identify, neutralize, or otherwise circumvent the normatively disabling effects of one's past and present environments. As I see it, the real threat of biological and social determinism lies in forgetting that we are, all of us, *still being determined* by our environments and *still learning* from each other.

My project is thus very much in line with Manuel Vargas' and Victoria McGeer's recent attempts to revive *forward-looking* theories of responsibility. Broadly speaking, a forward-looking theory of responsibility is one which argues that practices like praise and blame are justifiable because they tend to beneficially shape the way we think and act (Schlick 1966, Smart 1961).[1] Forward-looking accounts of responsibility have this major advantage: they explain why agents can be justifiably held accountable for harmful thought and behavior in a way that doesn't require ultimate control or radical freedom. Even where an agent's thought and behavior can be causally traced to factors which were outside of her control—such as chancy events, damaging formative experiences, the laws of the universe, or simply the past—she may still be justifiably held accountable inasmuch as doing so would augment or scaffold her capacity to think and act less wrongly. What matters is not whether things could have been otherwise, or whether alternatives were "genuinely" open to the agent, but is rather what might still be possible if we respond to one another appropriately.

Forward-looking theories of this *agency-cultivation* variety emphasize a certain developmental and historically contingent picture of responsible agency. They suggest that our status as responsible agents rises or falls with our continued willingness to treat one another *as* responsible agents through a host of agency-cultivating social practices. Responsibility is something we can lose, and it is something we are all the more likely to lose if we fail to

[1] Because these accounts see the justification of said practices as lying in their effects, they may also be aptly characterized as *consequentialist* accounts of responsibility, although the reasons for embracing a forward-looking account need not stem from a prior commitment to consequentialism, and it does not entail a commitment to consequentialism in other domains.

appreciate its socially contingent nature—if we mistake our freedom for something inalienable or fundamental to our being. Daniel Dennett puts the idea like this:

> We live our lives against a background of facts, some of them variable and some of them rock solid. Some of the stability comes from fundamental physical facts: The law of gravity will never let us down […], and we can rely on the speed of light staying constant in all our endeavors. Some of the stability comes from even more fundamental, *meta*physical facts: 2 + 2 will always add up to 4, the Pythagorean theorem will hold, and if A = B, whatever is true of A is true of B and vice versa. The idea that we have free will is another background condition for our whole way of thinking about our lives. We count on it; we count on people "having free will" the same way we count on them falling when pushed off cliffs and needing food and water to live, but it is neither a metaphysical background condition nor a fundamental physical condition. Free will is like the air we breathe, and it is present almost everywhere we want to go, but it is not only not eternal, it evolved, and is still evolving. […] The atmosphere of free will is another sort of environment. It is the enveloping, enabling, life-shaping, *conceptual* atmosphere of intentional action, planning and hoping and promising—and blaming, resenting, punishing, and honoring. We all grow up in this conceptual atmosphere, and we learn to conduct our lives in the terms it provides. It *appears to be* a stable and ahistorical construct, as eternal and unchanging as arithmetic, but it is not. It evolved as a recent product of human interactions, and some of the sorts of human activity it first made possible on this planet may also threaten to disrupt its future stability, or even hasten its demise. Our planet's atmosphere is not guaranteed to last forever, and neither is our free will. (Dennett 2003, pp. 9-10)

Once this social-developmental picture is clearly at hand, its indispensability to how we theorize about responsibility becomes difficult to refuse. Probably most theories of responsibility, even the most radical libertarian and hard incompatibilist stripes, accept that humans are developmental creatures who learn from each other, and that this learning helps humans to think and behave more skillfully and ethically. Even if we really are capable of beginning causal chains, and even if we possess some measure of innate moral knowledge, this does not change the fact that humans will exercise their freedom and wield their moral knowledge in ways that are highly sensitive to circumstance and education. Theorists should be wary not to let a fixation on *necessity* and *essentiality* draw their attention away from something so obvious. And while I too will make my own sort of wager to the effect that learning is necessary for a certain kind of responsibility, and an essential piece of what it is to be a certain kind of responsible agent, I am more interested in showing what humans are altogether capable of, above and beyond the bare

necessities, and how these capacities ground distinct forms of normative appraisal and norm-governed relationships.

Forward-looking theories of responsibility do have their challenges, however. For one, they are in certain respects at odds with the ways we ordinarily think about responsibility. Except perhaps in our roles as parents and educators, we don't always think of our accountability practices in this way—as efforts to change the way people think and act. People's minds may change as a result of our blaming them, but these changes are not always what we *intend*, nor do they seem to be what *justifies* our blame. We would blame someone all the same, and we would think ourselves justified in blaming them, even if they were dead and gone, or if they were simply too stubborn, or too corrupt, to see the error of their ways. To paraphrase Strawson, in reacting to someone in these ways, we do not think of her as an object of policy or manipulation, someone to be brought into line with carrots and sticks.

Moreover, what I earlier called an advantage of forward-looking theories is arguably a *disadvantage*. After all, if a person's thinking and behavior is ultimately traceable to factors outside of her control, shouldn't this count as a legitimate *excuse?* And if something like this is true of all human activity, wouldn't we thus be *exempt* from practices of responsibility altogether? Even if we are ordinarily disposed to treat one another as responsible agents, conserving that standing practice is not an advantage if it is unjustifiable. The kind of conservation that is needed is not merely a satisfactory descriptive account of how responsibility practices work and the capacities of human agency that make them work. It must also tell us why these practices and capacities are something worth keeping—why we are justified in sustaining the kind of social atmosphere in which those capacities evolved and continue to evolve.

I will have a lot more to say about these two challenges, but for the time being let this promissory note suffice: while I agree that we do not tend to think of responsibility in this way, I argue that the justifiability of practices like praise and blame does partly depend on whether the agent can learn from them. There are certainly instances in which agents are excused or even exempted because of factors they could not have avoided or altered, but there are also many instances in which that sort of tracing does not undermine their eligibility to be held accountable and, more broadly, to participate in relationships of accountability. Whether and how we sustain an atmosphere of responsibility are also questions of whether and how we might tweak and adapt that atmosphere *so as to be* more sustainable, given what we are coming to learn about our own

nature and origins. In certain respects, we may need to learn to think differently, or at least learn to appreciate that the way we ordinarily think about these things is at best superficially correct, and that there is much more going on beneath the surface.

Besides answering this basic question of what makes humans responsible for what they do—the subject matter of papers one, two, and three—my project is also aimed at advancing a number of theses about interpersonal relationships and personhood, particularly as these concepts relate to and can be informed by empirical accounts of human agency and psychology. In papers four and five I consider how certain descriptive features of human agency and psychology may shed light on emerging social AI technologies and substance use disorder, respectively. I attend to these issues in part because of their close connection to questions of responsibility. For one, it is contested whether individuals with addiction are responsible for the things they do as a result of their addiction, and much of this controversy turns on the question of whether such individuals either choose to act in the ways they do or are instead compelled by an underlying neurobiological condition. My particular point of concern in paper five will be with the role that *personhood* plays in addiction and the impact this has on how we understand individuals' agency with respect to addiction, be it in the context of treatment or in our ordinary relationships with such persons. As for social AI, there is also debate as to whether such "artificial agents" are (or could be) *responsible* agents. But rather than attend to this fraught metaphysical question, my goal in paper four will be to consider whether and how humans should relate to social AI, given that the latter are designed with the express intention of participating in interpersonal relationships with humans, and given that ordinary interpersonal relationships are structured by the psychology and practices of norms and responsibility.

Glossary of Terms

The dissertation will in various instances appeal to a number of definitions and conditions. The hope is that these definitions and conditions are put in such general and uncontroversial terms that they can serve as a basis upon which to advance some more specific and less widely accepted claims.

1. Normative responses

A theory of responsibility centers on this ordinary social fact: that some human activity tends to elicit and even appears to warrant a range of *normative responses* from other humans. Praise, blame, gratitude, resentment, distrust, reward, punishment, distancing, confrontation, forgiveness—the list goes on. Not only do humans experience strong motivations to respond to others' behavior in these ways; they also tend to think that they have good reasons for doing so.

In calling these "normative responses" I mean to signal that they are social responses which concern the proverbial "dos and don'ts" of human life—the things one should and shouldn't do, feel, think, or say. *"Don't drink and drive." "Say, 'Please' and 'Thank you'." "Don't be racist." "Tip the waiter." "Be true to your word."* That is to say, they are responses to the way humans observe or fail to observe certain social or moral rules—or norms—which govern what is appropriate, allowed, required, or forbidden in different situations and relationships and for different members of a community. Many norms thus pick out what in ethics is called an obligation or duty—something which the agent who is subject to the norm ought to do. Though I do want to flag that in a given society there may be norms which are ethically problematic. In other words, there is a descriptive question about which norms (duties, obligations, expectations) are observed and enforced in a given social context, and this question stands apart from the normative question about which norms are *legitimate* or *justified.* Also, the specific contents of these norms often varies from one culture or social context to the next, but the existence of some system of normative expectations and reactions appears to be culturally universal.

Normative responses are *evaluative* in an important way. They represent their targets as meeting, exceeding, or violating some normative standard or rule. Differently stated, they are forms of *normative appraisal.* So, for example, whether someone is praiseworthy for an action depends on whether the person is legitimately subject to some norm and whether the action in question met or exceeded that norm. Similarly, whether someone is blameworthy for an action depends on whether she is legitimately subject to some norm and whether the action violated that norm.

2. Responsibility

In common usage, the English words "responsibility" and "responsible" have a number of different meanings. No one genuinely *blames* explosions for the harms they cause—they blame the people who denoted the bomb—but people certainly do say things like, "The explosion was responsible for over 200 deaths." That is because "X is responsible for Y" sometimes just means that X caused Y. Understood in this way, responsibility is a purely descriptive matter—it describes one thing or event, Y, as depending for its existence or occurrence on some other thing or event, X.

The sense of responsibility which is at issue in a theory of responsibility, by contrast, is not purely descriptive. Here "X is responsible for Y" means that X is *eligible for a response*—the normative responses glossed in the previous section. To say that X is eligible for these responses is to say that responding to X in one of these ways would be justifiable, at least in principle.

By analogy, consider what it means to say that something is potable. "The water in Flint, Michigan, is not potable" means that the water there is not safe to drink. It does not mean that the water is not *drinkable*. The problem is not that the water in Flint cannot be drunk. It is that it *should not* be drunk. And the reasons there are not to drink the water have do to with (a) descriptive facts about the water's contents and about the effects those contents have on the human body and (b) things one is presumed to value—i.e., being in good health, or not being in bad health. This constellation of facts and values is "reasons-giving" in the broadest possible sense: they come together to support claims about what one should or should not do, such as: "You shouldn't drink the water."

The concept of potability is thus both descriptive and normative. It is what Bernard Williams calls a "thick concept". Similarly, "X is not responsible for Y" does not mean that X cannot be praised or blamed for Y, in the sense that it is impossible to do so. Rather, it means that X *should not* be praised or blamed for Y—that it would be inappropriate or unjustified or wrong. Ascribing responsibility to someone for her behavior, then, does not just describe her as bearing some relationship to this behavior, e.g., bearing a causal relationship to the movements of her body. It also says that the manner in which she is related to her conduct is sufficient to make her an appropriate target of normative responses.

Following T. M. Scanlon (1998) and Angela Smith (2015), consider the following preliminary definition of the relevant sense in which an agent can be responsible for something:

> Responsible (preliminary): *An agent, S, is responsible for something, X, just in case X can be attributed to S in a way that makes S eligible for normative responses on the basis of X [e.g., praised or blamed for X].*

There are two things I want to bring to the fore about this definition. First, although it specifies that the manner in which something must be attributed to the agent is one which makes the agent eligible for normative responses, it does not say what the requisite manner of attribution exactly amounts to. That is, it leaves open just what sort of attribution is required—mere causal attribution, for example, or something more involved. Second, although it does not explicitly say so, whether something can be attributed to an agent in the manner required for normative responses depends on *the kind of agent it is* (and whether it is an agent at all). That is to say, depending on what the requisite manner of attribution amounts to, only certain kinds of things will qualify. If the requisite manner of attribution exceeds mere causal attribution, then the being in question will need to have capacities that go beyond the capacity to cause things to happen.

3. Responsible agency

So, in order to be responsible for what it does, the being in question must be a specific kind of agent. Manuel Vargas says about this idea that "judging that someone is responsible involves acceptance of a (at least implicit) judgment that the evaluated agent is a special kind of agent, what I will call *a responsible agent.* ... So, we need a theory that recognizes a restriction on the kinds of agents that are properly subject to moral praise and blame" (Vargas 2013, p. 110). Let this very general notion of responsible agency be defined as follows:

> Responsible agent: *A being, B, is a responsible agent just in case it is, in principle, an eligible target of normative responses.*

Arguably, the preliminary definition of responsibility given in the previous section obscures this aspect of responsibility, or else leaves it at an implicit level. That definition is, after all, restricted in scope to *agents*, and so it rules out or is at least silent on the possibility that explosions or tornados could be responsible for what they do. But surely a desideratum of any satisfactory theory of responsibility is that it specify just what sort of being—e.g., all agents, or only specific

kinds—is "properly subject" to normative responses and also explain *why* that sort of being qualifies whereas others do not.

As an anticipatory note, much of what I will be concerned with in the present project is this aspect of responsibility: what responsible agency consists in and why. But there are important connections between all of these elements. Whether a given being is a responsible agent will depend on whether it is the sort of being which could have things attributed to it in the way required for normative responses, and vice versa, and both will depend on what makes something a normative response. So, in concerning myself primarily with the conditions of responsible agency, my discussion will regularly appeal and attend to considerations having to do with attribution and with the nature of normative responses.

To draw out this desideratum more explicitly, let the definition from the previous section be revised in the following way:

> Responsible (revised): *A being, B, is responsible for something, X, just in case (1) B is a responsible agent and (2) X can be attributed to B in a way that makes B eligible for normative responses on the basis of X [e.g., praised or blamed for X].*

4. Being responsible vs. having a responsibility

Sometimes the word "responsibility" is used as a synonym for "duty," "obligation," or "expectation." To say that the gardener *has a responsibility* to water the garden, for example, is to say that the gardener, plausibly in virtue of the role she occupies, *ought* to water the garden, is under an obligation to do so, is reasonably expected to do so, etc. The same idea might even be expressed as, "The gardener is responsible for watering the garden," meaning that this is something the gardener is tasked with doing—it is something required or expected of her *qua* gardener.

This usage is liable to introduce a bit of confusion, and so I will avoid it to the best of my ability. But I do want to draw some attention to it, because when we consider the question of whether a being is responsible for what it does, it is sometimes easy to slip into thinking about whether the being in question *is legitimately subject to some norm*. If asked whether Joaquin is responsible for refusing to give money to a homeless person, someone might respond by saying that Joaquin is under no obligation to give money to homeless people—he *has* no such responsibility—and therefore he *is* not responsible for refusing to do so. But this inference is

invalid. Even if it is true that Joaquin is under no such obligation, he may still *be* responsible for his decision, so long as he is a responsible agent and his decision can be attributed to him in the right way. Likewise, if while sitting alone at my desk I voluntarily raise my own hand, I am probably responsible for this action, even though I am under no obligation to raise or not raise my hand and, more generally, even though I am not legitimately subject to any hand-raising norm (at least in this context). Summarily stated, it is possible to be responsible for things which are normatively neutral.

This is important because it brings into clearer focus just what is meant by saying that someone is responsible for something, understood in terms of the agent's eligibility for a normative response. It just means that a normative response *would* be justifiable, *should it turn out* that the thing in question actually meets, exceeds, or violates a norm to which the agent is legitimately subject. It does not carry any commitment as to *which* normative response, if any, is called for. If Joaquin is responsible for his decision not to give money to homeless people, then if he is in fact under an obligation to do so, it would be justifiable to blame him for it. Joaquin's blameworthiness thus depends on two separate conditions:

A. Joaquin is responsible for his decision (he is the right kind of being and his decision is attributable to him in the right way).	B. Joaquin is legitimately subject to a norm (e.g., an obligation to give money to the homeless) which his decision has violated.

By contrast, if Joaquin were under an obligation to *avoid* giving money to the homeless, then he would be (or at least could be) praiseworthy for what he does. But in either case the justifiability of responding to Joaquin with a specific normative response depends on facts that go beyond his being responsible for his decision—it depends, namely, on what is legitimately required or expected of Joaquin.[2]

Generally speaking, a theory of responsibility is concerned only with the condition picked out in (A). Whether and why Joaquin has a particular obligation to give money to the homeless, and in general questions about which norms humans are legitimately subject to, be it

[2] Vargas puts the idea like this: "our assessments of moral praiseworthiness and blameworthiness appear to be parasitic on our assessments of what morality requires of us. Moralized blame presumes that we have done wrong. Moralized praise presumes that we have done right" (Vargas 2013, p. 111).

universally or in specific cases, is the object of inquiry of normative ethics. To be sure, questions about responsibility are situated in a theoretical space that overlaps with normative ethics, but a theory of responsibility can take for granted that normative ethics is not a bankrupt enterprise—that *some* such legitimate norms exist. Obviously, if humans are not legitimately subject to norms at all, then the significance of the question of whether they are nonetheless *responsible* for what they do would be rather mysterious.

That being said, and as a second anticipatory note, I believe there is something deeply objectionable about approaches to responsibility which draw the line between these two projects too boldly. As I'll discuss in the first paper, whether a being is *capable* of being legitimately subject to norms—whether it is the kind of being which even could be obligated or reasonably expected to do specific things—is relevant for determining whether that being qualifies for responsible agency. Certain theories of responsibility, and usually those which deny that either *freedom* or *control* is a necessary condition of responsibility, come dangerously close, in my view, to rejecting this important connection. At the very least, they obfuscate it by treating considerations about freedom and control as though they pertain only to the question of whether an agent has some *particular* obligation or is *excused* from some such obligation. As I hope to make clear, even the basic conditions of responsibility include a control or freedom-relevant requirement in virtue of the role these capacities play in responsible agency.

5. Excuses and exemptions

To close this section, I want to briefly sketch, rather than define, two conditions which speak against holding someone responsible: *excuses* and *exemptions.* Although Strawson does not call them by these names, the distinction between excusing conditions and exempting conditions is present already in his well-known essay, "Freedom and Resentment" (1962). There Strawson contrasts two kinds of considerations which tend to modify or suspend the *reactive attitudes* we feel toward an agent, where we can understand reactive attitudes as falling under the broader umbrella of normative responses.

As examples of excuses Strawson mentions things like accidents, ignorance, compulsion, coercion, duress, and manipulation. In cases such as these, Strawson argues, it is not that the agent is an altogether inappropriate target of the reactive attitudes (i.e., it is not that the agent is not a responsible agent). Rather, it is that the specific action in question fails to qualify for a

specific response. Whether it is justifiable to blame an agent for some putative wrongdoing, for example, depends on whether the agent's circumstances (at the time of action or earlier in her life) are such that we cannot reasonably expect her to have done otherwise. Perhaps she lacked an adequate opportunity to avoid the wrongdoing, as when an agent's will is "overborn" by manipulative or coercive circumstances. Perhaps she "wasn't herself"—she was under duress or was suffering from some temporary malady or condition. Perhaps what she did was actually the right call given the wider array of considerations bearing on her decision. Or perhaps she lacked some crucial piece of information, such as that her action would have a disastrous consequence, as well as an adequate opportunity to acquire that information. These are just some plausible candidates, and are not meant to exhaust the possibilities. They do well enough to illustrate the sense in which an agent who is nonetheless a responsible agent could be excused for thinking, saying, or doing something that ordinarily speaking she is obligated or rightly expected to avoid. Or, stated somewhat differently, they illustrate the sense in which a norm which is nonetheless legitimate in general may be sensitive to extenuating circumstances and overriding considerations. Underscoring this idea, Strawson says that they "do not suggest that the agent is in any way an inappropriate object of that kind of demand for goodwill or regard which is reflected in our ordinary reactive attitudes. They suggest instead that the fact of injury was quite consistent with the agent's attitude and intentions being just what we demand they should be" (Strawson 1962, 8).

By contrast, an agent is *exempt* just in case she is an altogether inappropriate target of normative responses (i.e., is not a responsible agent). Here Strawson's examples include chronic psychological disorders, as well as moral underdevelopment and poor formative circumstances. There is, to be sure, much controversy about what it takes for a person's upbringing or psychological condition to wholly disqualify her from normative responses. But there is no controversy as to whether *some* beings simply fall short of the kind of agency required, and most of us would probably agree that even some humans, such as those suffering from serious cognitive disabilities, are also altogether exempt from being held responsible. In Strawson's words, we may respond to such an agent in a wide variety of ways, but our responses "cannot include the range of reactive feelings and attitudes which belong to involvement or participation with others in inter-personal human relationships," for which reason he called these *participant* reactive attitudes (Strawson 1962, pp. 9-10). That is to say, our responses must be sensitive to the

fact that not just everything and not just everyone can be a participant in normatively structured relationships—the kind in which agents are legitimately *governed* by shared normative principles or standards.

Following Vargas, we can summarize all of this by saying that "exemptions track the nature of the agent" and they arise "because of failures to be the right kind of agent" (Vargas 2013, pp. 112-113). By contrast, excuses "arise in cases where the action was not of the right sort" (i.e., is not something for which the agent can justifiably be held responsible), such that an agent who is nonetheless a responsible agent "might still fail to be praiseworthy or blameworthy because of the kind of action it was, or the manner in which the action as produced" (ibid.).

Summary of Papers

Paper 1 – The Capacity to Have Responsibilities

The first paper of the dissertation is best understood as a forward-looking response to the broader debate between control-based theories of responsibility (such as reasons-responsiveness theories) and theories which reject the control condition (such as deep self and answerability theories). Forward-looking theories of responsibility come into this debate in a particularly vulnerable position: not only do they need to show that control is a requirement for responsibility, but they must also show why the requisite form of control should have anything to do with learning. My goal in the first paper is to satisfy these desiderata.

The paper is divided into two parts. In the first part I argue that, in order for an agent to be responsible for what she does, she must be at least generically capable of *having* responsibilities. Differently stated, she must be the sort of agent who can be legitimately subject to norms. As I understand it, this is the point Strawson is making when he says that, in order to be eligible for the participant reactive attitudes, an agent must be an "appropriate object" of the basic demand for goodwill. This is, perhaps, an uncontroversial claim: if we couldn't even *have* obligations, then we certainly couldn't be praiseworthy or blameworthy for how well we fare with respect to those obligations. Still, there are some theories which I believe fail to appreciate the constraint this places on a satisfactory account of responsibility. Both the deep self and the answerability approaches to responsibility—what I call *self-direction* theories—offer relatively elegant accounts of responsibility. They claim that the agent is responsible for what she does

inasmuch as it *expresses* the agent's self or *reflects* her judgment. But that elegancy is due in large part to the way self-direction theories reject both *control* and related *competency* requirements for responsibility. If responsible agency is merely a matter of authorship, then the agent's capacity to do otherwise and even her responsiveness to normative considerations are beside the point. I think this is a mistake. But in order to best illustrate my reasons for thinking so, I first need to establish a desideratum of any satisfactory theory of responsibility: namely, that it account for the agent's capacity to have substantive normative obligations (or, be legitimately subject to norms). That is the aim of part one.

In part two I argue that the capacity to have substantive normative obligations is a dynamic form of reasons-responsiveness, one which is characterized by learning from corrective normative feedback, such as that which is communicated in praise and blame. I illustrate this point by contrasting a pair of agents who are capable of the same activity (playing chess), and who act in ways that reflect their judgment, but who differ with respect to the manner in which they may be *normatively appraised.* Appraisal looms large in the answerability theory, because the theory understands responsibility as the agent's eligibility to be appraised (e.g., criticized) for how well or poorly she conforms to normative standards. Following Strawson's distinction between the objective and the participant standpoints, I argue that some agents are open to being normatively appraised in a restricted sense—we can evaluate them from a technical standpoint, observing that they exceed, meet, or violate the norms governing the activity in question. But they are not open to being normatively appraised in the manner that is at issue in normative responses like praise and blame. These are responses which communicate not merely a technical evaluation of excellence or deficiency, but a statement to the effect that the agent has done well or poorly *by some norm she is legitimately subject to* (one she has an obligation to conform to). As such, they represent their targets as relatively sophisticated agents—ones which may reasonably be expected to conform to normative standards, even and perhaps especially in the event that they fall short of those standards. If ought implies can, then the agent in question must be capable of more than rational self-direction. She must also be capable of learning to bring herself more closely in line with the norms to which she is legitimately subject, which I argue requires a specific type of responsiveness to corrective feedback. As I discuss, this dynamic form of reasons-responsiveness has roots in Gilbert Ryle's notion of an *intelligent capacity*, which Victoria McGeer has since expanded upon.

The paper thus argues that, if indeed we understand responsibility as eligibility for the kind of normative appraisal that is at issue in things like praise and blame, and if that kind of appraisal represents its target as capable of living up to norms, then responsible agents must be those who can bring themselves into conformity with norms. This is an agent who not only can be evaluated against normative principles and standards, one who happens to conform or happens not to conform to their prescriptions, but who can also be *held* to those principles and standards in the event that she violates or falls short of them. Her conformity with norms is "up to her" in a way that it clearly is not for a merely self-directed agent. Without this further capacity, the agent would be ineligible for normative responses like praise and blame—she would fail to be a *responsible agent*—even if she is still the author of her own conduct.

Paper 2 – Teaching an Old Dog New Tricks: Intuition, Reason, and Responsibility

The second paper of the dissertation serves as something of an application of the points raised in the first. Ultimately it argues that, because a human's capacity to live up to norms depends on social corrective feedback like praise and blame, it can be justifiable to hold humans responsible for things like implicit biases, inasmuch as these accountability practices augment the person's capacity to mitigate or handle her biases appropriately.

The paper is presented as an attempt to address a potential worry for control-based theories of responsibility having to do with the more "automatic" and even "unconscious" features of human agency. It argues that control-based theories of responsibility can explain why humans are sometimes responsible for what they do as a result of affective and intuitive processes (such as behavior resulting from problematic implicit biases) if those theories avail themselves to certain forward-looking resources. At first glance, control-based theories appear to be at odds with the idea that humans could be responsible for things they do as a result of processes of which they are not aware and over which they do not have direct control—i.e., by way of explicit deliberation. Arguably, this tension between control and automaticity is part of what is motivating some theories to eschew the control requirement altogether in favor of less agentially demanding answerability or self-expression requirements. However, as I discuss, even if we accept that these processes are largely insensitive to the agent's explicit judgment—her conscious and deliberate response to reasons—they may still be under the agent's control in less direct ways having to do wither her environment and especially her social environment. I argue

that control-based theories can explain responsibility for things like implicit bias if they are revised along agency-cultivation and forward-looking lines.

The paper begins by presenting the reasons-responsiveness approach to responsibility (Section 2) and the challenge presented to this approach by an emerging picture of human psychology (Section 3). According to this highly influential control-based approach, 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 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 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 pieces of environmental scaffolding by which human beings are attuned to reasons over time.

The paper proceeds by considering two "steps in the right direction" toward a revised reasons-responsiveness theory, but which are insufficient on their own to answer the challenge presented by the dual-process model. The first step (Section 4) is to argue that even affective and intuitive processes can be reasons-responsive (Railton 2014, 2017), 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. The second step (Section 5) 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, which I discuss in Section 6, 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 constitute 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 depends. I argue that the resulting account can explain why human beings are sometimes blameworthy even for things which result from unconscious, intuitive processes. 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. And it is capable of learning precisely by *holding each other accountable* for our reasons-responsiveness failures.

Paper 3 – Deep Selves are Just Special Kinds of Reasons-Responders

The third paper of the dissertation departs somewhat from the earlier two in that it does not advance a specifically forward-looking thesis. Like the second paper, however, it responds to an objection against control-based theories of responsibility, this time having to do with the role that *flexibility* plays in standard reasons-responsiveness theories (Sripada 2017). And like the first paper, part of the motivation in responding to this objection is to push back against the broader self-direction approach to responsibility, and to do so in a way that vindicates the control requirement.

In the paper I argue that deep self theories are ultimately a *kind* of reasons-responsiveness theory. Chandra Sripada (2015) has already persuasively argues that the two theories bear a close relationship to one another, but Sripada maintains that they are different in one important respect:

that reasons-responsiveness theories consider *flexible control* to be a necessary condition of moral responsibility, whereas deep self theories do not. Using resources from McKenna (2013), I argue that Sripada has misunderstood the function that flexibility plays in reasons-responsiveness theories, which leads him to disagree with John Fischer and Mark Ravizza's (1998) account for the wrong reasons.

In Section One and Section Two I summarize the two theories under consideration: Fischer and Ravizza's reasons-responsiveness theory and Sripada's self-expression theory. Section three reviews Sripada's reasons for thinking that deep self and reasons-responsiveness theories are difficult to disentangle, as well as his reason for thinking that an importance difference nonetheless remains between the two. In Section Four I consider the cases which Sripada believes illustrate this difference, and which he uses to argue in favor of the deep self approach. In Section Five I object to Sripada's conclusion on the grounds that he has misunderstood the flexibility requirement, and that a proper understanding will reveal that Sripada's objection against control is actually an objection against Fischer and Ravizza's notion of a *mechanism of action.*

As I briefly take up in the paper's concluding section, I believe that the real point of substantial disagreement between specific deep self theories and specific reasons-responsiveness theories lies in *which agential features are responsibility-relevant.* While I do not explore this point further in the paper, I believe that deep self and similar answerability theories fail to identify the right agential features, and for reasons similar to those given in paper one. Self-direction theories are admittedly more parsimonious than control-based theories, but that simplicity comes at the cost of ignoring the importance of an agent's capacity to bring herself up to normative standards, rather than merely act well or poorly by those standards in accordance with her own judgment or cares. A responsible agent is one who not only acts on her own judgments, or acts in ways that are motivated by what she cares about, but is also capable of *changing* her judgments and motivations in response to corrective feedback.

Paper 4 – A Willingness to be Vulnerable: Norm Psychology and Human-AI Interactions

The fourth paper of the dissertation raises and then addresses a specific worry about relationships between humans and certain artificial agents called "social robots". I begin with the observation that social robots are designed to be welcomed into roles and relationships which are

characterized by a high degree of *vulnerability*: caregiving and assisted living roles, educational roles, and platonic as well as romantic and sexual partnerships. The vulnerable nature of these roles and relationships is underscored by the selectivity with which we ordinarily welcome someone into them. In ordinary human relationships, when we put ourselves or a loved one under someone's care, or when we open up to a friend or a lover, it is typically because we think we can *trust* that they will be responsive to our or our loved one's expectations about how we should be treated. The guiding question of the paper is this: Should we welcome social robots into these roles and relationships of vulnerability, and if so, on what conditions? By posing the question in this way, I follow what Dorna Behdadi and Christian Munthe (2020) call a "normative approach" to artificial agency. Instead of debating about the necessary and sufficient conditions for moral, social, or responsible agency, I will attend more directly to the practical question of whether we *should* welcome social robots into roles and relationships which characteristically involve a high degree of bodily, psychological, and emotional vulnerability on the part of human users. I argue that, as in the human case, our willingness to welcome social robots into these spaces, roles, and relationships should be conditioned on their capacity to understand and live up to *social norms.*

In Section One I reflect on the importance humans attach to certain social expectations in their ordinary interpersonal relationships, and I give reasons for thinking that we should look to these relationships as models for the expectations we would have of social robots. Section Two expands on the human tendency to attach significance to interpersonal expectations by giving a broad overview of the psychological capacity which enables humans to detect, internalize, and be motivated by these expectations—or, *social norms*. There I emphasize the role that a variety of *normative responses* play in communicating and enforcing these expectations. Then in Section Three I argue that, just as in the human case, we should expect more from social robots than mere rule following. What is required is a higher-order capacity to *update* behavior in response to the corrective feedback which users provide through normative responses like praise and blame. On the picture I present, our willingness to be vulnerable with social robots should be conditioned on a specific type of social-normative learning.

Paper 5 – Addiction is Not Diminished Personhood: The Role of Dependencies in Self-Governance

The dissertation's fifth and final paper is a critical response to a commonplace view of addiction as a form of diminished capacity to self-govern. Don Ross has recently taken up this broader picture and defended a specific variety of the view according to which addiction is a form of diminished personhood, where personhood is understood in terms of the agent's capacity to govern herself according to personal rules. As I indicate, however, there is more to addiction than the familiar phenomenon in which an agent acts contrary to her own best judgment. If that were the sole explanatory target of a theory of addiction, then it would only be natural for the theory to place emphasis on a process, neurobiological or otherwise, whereby the individual's capacity for self-control or self-governance is overpowered, hijacked, bypassed, or diminished. But while this phenomenon is no doubt central to many cases of addiction, sometimes addiction is a product of the *way* an individual self-governs, rather than her incapacity to do so. That is to say, sometimes the reason an individual with addiction engages in addictive behavior is not that she has failed to govern herself according to her better judgment, but is rather that she has successfully governed herself according to a particular vision of who she is, what she values, and what she is capable of.

In Section One I present Ross' *Diminished Personhood Account of Addiction,* and I discuss why it is so compelling, given that it embraces both the neurobiological and the choice-theoretic aspects of addiction. In short, Ross argues that individuals with addiction come to see themselves as incapable of following their own personal rules, due to neuroadaptations that result from addictive behavior and undermine their capacity to enforce those rules. Then in Section Two I show that individuals with addiction may interpret their relationship to their addiction rather differently. Namely, they may come to represent themselves as dependent upon the substance or activity in question in order to achieve the things they value. In this way, addictive behavior may come to have instrumental value that leads individuals to adopt personal rules *for,* rather than *against,* that behavior. In Section Three I conclude by addressing a number of objections.

The fourth and fifth papers do in a number of important respects "stand alone" from the broader theory of responsibility developed in the first three papers. In other respects, however, they provide a fuller view of the features of human agency which would support that theory at a

descriptive level. For example, the fourth paper is first and foremost a paper about what we should do in the face of emerging AI technologies and applications—whether and how we should relate to so-called "social robots"—but it also summarizes a general-level theory of norm psychology that highlights the importance of normative learning and feedback. Similarly, while the fifth paper argues for a specific descriptive claim about substance use disorder, it also gives the accounts of *personhood* and *self-governance* in which my broader theory of responsibility is situated. Arguably, the fourth paper stands in closer connection that theory, because it describes human relationships as depending on a measure of normative *dynamicity* between the individuals in those relationships. If this is right, then participation in those relationships requires more than a first-order capacity to follow the rules or norms of the relationship. It also requires a second-order capacity to update this first-order capacity in light of the way the relationship "plays out" or evolves over time—as when the individuals within the relationship change their own understanding of how they ought be treated and what is reasonable to expect. As for the fifth paper, while its connection to considerations of responsibility is less direct, it does have at least one such connection: in cases where addiction is at least partly a product or expression of the agent's self-governance, there is some reason to think that accountability practices would not be altogether inappropriate, so long as they occur between properly situated agents (agents who stand in specific types of relationships with the individual), and so long as those practices are modified in line with what is reasonable to expect of the individual (how much she can reasonably be expected to change).

THE CAPACITY TO HAVE RESPONSIBILITIES

Abstract. Being responsible *for* something is distinct from *having* a responsibility, or an obligation, to do something in particular. In this paper I argue that responsibility in the former sense requires the capacity to have responsibilities in the latter sense, and that this may have been why P. F. Strawson (1962) defined responsible agency in terms of being an "appropriate object" of the demand for goodwill—i.e., an agent with the capacity to understand and live up to interpersonal norms.

The answerability theory of responsibility (Scanlon 1998; Smith 2015), which exemplifies an increasingly popular self-direction approach to responsibility, rejects this Strawsonian idea and, along with it, various control-based conceptions of responsibility. But I maintain that the capacity to have responsibilities is a necessary condition of responsibility and, moreover, involves a more sophisticated form of agency than rational self-direction. Namely, it requires the dynamic form of reasons-responsiveness exhibited by what Gilbert Ryle (1949) calls an "intelligent capacity."

Introduction

An increasingly popular view of responsibility is that it consists in being the causal or justificatory *source* of one's thought and behavior.[3] For example, Chandra Sripada's (2016) self-expression account, a variant of the deep self view of responsibility, holds that an agent is responsible for something just in case it was motivated by certain conative states called "cares". And the answerability theory, as articulated and defended by T. M. Scanlon (1998) and Angela Smith (2015), holds that agents are responsible for something just in case it was governed by or reflects the agent's judgment. Understood in this way, responsible agency is a form of *self-direction* or *self-disclosure*. Actions and attitudes are self-directed, or self-disclosive, just in case they are caused by or reflect some feature of the person's agency with which she can be identified, such as her desires, her reasons, or a distinguished subset of these.

I believe something is missing from this picture of responsibility. Humans are not just self-directed or self-disclosive. They are also constantly self-updating, dynamic learners who are especially attuned to the input they receive from other humans. We do not just do; we also

[3] For a discussion of *source freedom* and *leeway freedom* as different conceptions of the freedom or control condition of responsibility, see McKenna (2013), p. 152.

change what we do and why we do it as a function of the feedback we receive from our environments, and especially from our social environments. If this sort of agency is not distinctive of humans, it is at least our forte,[4] and is therefore a plausible place to go looking for the capacities which make us responsible agents, given we expect these to be capacities possessed by most humans but not most nonhuman animals.

In this paper I focus on the answerability theory as an exemplar of the broader self-direction and self-disclosure approach, and I argue that it gives an inadequate account of responsibility for two, interrelated reasons: (1) in order to be responsible for anything she does, an agent must be *capable of having responsibilities*, and (2) the capacity to have responsibilities requires a dynamic form of agency which well exceeds self-direction or self-disclosure. My argument thus builds out of Scanlon and Smith's distinction between *basic* responsibility and *substantive* responsibility—between being responsible *for* something, on the one hand, and *having* a responsibility, on the other. In Section One, I review this distinction and identify a difficulty it seems to raise for theories, like P. F. Strawson's (1962) reactive attitudes theory, which maintain that agents must be capable of understanding and living up to normative demands in order to be responsible for what they do. I believe that difficulty can be resolved, however, once we appreciate how closely related these two senses of responsibility actually are. In Section Two I argue that, in order to be responsible for anything that she does, the agent must also be, in some generic sense, an appropriate object of normative demands. That is, she must be the sort of agent who is capable of having a substantive responsibility to conform to interpersonal norms. For otherwise she could never be eligible (even just in principle) for praise and blame.

The answerability theory could accommodate this conclusion by maintaining that rational self-direction (answerability) is sufficient for specifically moral responsibilities, and its defenders do appear to embrace this idea. To address this, in Section Three I argue that the capacity to have substantive responsibilities of any kind requires more than answerability. I

[4] Recent developments in evolutionary theory suggest that the success and distinctiveness of the human species is due not so much to our intelligence or "big brains," but rather to our hypertrophied capacity for cumulative culture and social learning. See Boyd (2018) and Richerson & Boyd (2005) for a comprehensive account along these lines. Part of the motivation behind this paper is to articulate the formal framework for a theory of responsibility that is centered on this capacity for social learning.

argue that, from the mere fact that a creature has acted in a way that reflects its judgment, nothing at all follows about whether that creature can also be *obligated* to conform its judgments to the prescriptions of norms. What makes humans capable of having substantive responsibilities must therefore consist in something else. As I'll try to show, this is just a specific application of the general point, that there is nothing about merely having an agential capacity that makes the possessor of that capacity subject to the norms associated with it. What is required is a specific *kind* of capacity: one that exhibits a dynamic form of reasons-responsiveness—one, stated simply, which is characterized by *learning.*

Section 1 – Basic and Substantive Responsibility

Philosophers who work on responsibility tend to agree that a responsible agent is one who is eligible for normative responses like praise and blame.[5] The conditions of this eligibility, however, are subject to much debate. One particularly contested issue is whether, for at least some of these responses, the agent must display some measure of normative competency. Most well known, perhaps, is the challenge presented by *psychopaths*—agents who, per philosophical stipulation, act for their own reasons and from their own desires, but are incapable of responding to specifically moral reasons.[6] So characterized, they are agents who not only do not, but *cannot* see another's suffering as a reason to avoid causing it. One reason to think that psychopaths and similarly disabled agents are not responsible for what they do is that they are incapable of understanding and living up to the norms of collective life.[7] After all, in blaming someone we

[5] I have chosen the expression "normative responses" as opposed to "moral responses" (c.f. Smith 2015, p. 103), and I will in general speak of responsibility rather than *moral* responsibility, because I do not think of responsibility as being strictly a moral matter. Similarly, I will speak of praise and blame, rather than *moral* praise and *moral* blame, because I believe agents can be praiseworthy or blameworthy for how well they conform to all sorts of norms, rather than strictly moral norms (so long as they are legitimately subject to them).

[6] For a sample of that debate, see Levy (2007), Watson (2012), Scanlon (2013), Nelkin (2015), and Shoemaker (2015).

[7] Which is not to say that they cannot understand *that* others have normative expectations of them (see Smith 2015, p. 118, fn. 44). Rather, it is to say something along the following lines: that they are incapable of recognizing these expectations *as reasons-giving* in anything other than an instrumental sense. Also, to say that psychopaths are not responsible for what they do is not to say that it would be inappropriate or unjustified to prevent psychopaths from harming others, just as we would be justified in taking measures to prevent harms caused by nonhuman animals or natural disasters.

claim that the agent has done something wrong, i.e., something she has an obligation to avoid. But if ought implies can, then we are also implicitly claiming that the agent *could* have done as she should—that, unlike a psychopath, she had what it takes to judge and act in accordance with the obligation or norm in question.

However, others have argued that the *basic* conditions of responsibility do not include any such capacity or competency requirement. Rather, an agent's responsibility *for* what she does only requires that the action or attitude be related to the agent's *self* in the right sort of way—such as by reflecting her judgments, or by expressing her fundamental desires. On this sort of view, the fact that a given individual has had especially damaging formative experiences, or suffers from a disabling psychological condition, may be a reason to modulate or temper the *way* we hold her responsible, but it does not count as a reason for thinking the agent is not *eligible* to be held responsible.

The former position is most commonly associated with control-based conceptions of responsibility, such as John Fischer and Mark Ravizza's (1998) well known reasons-responsiveness theory, and it is also a part of pluralistic accounts that identify *accountability* as a distinct type of responsibility, such as Gary Watson's (2004) and David Shoemaker's (2015, 2011). But in this section I will be concerned mainly with its ties to Strawson's seminal essay, "Freedom and Resentment" (1962), in which he observed that whether we consider someone to be a responsible agent is sensitive to whether we believe the person is an "appropriate object" of the basic demand for goodwill. By reflecting on the importance Strawson attributed to this basic demand, and especially to the role he thought it played in *norm-governed relationships,* I hope to shed some light on the debate between control- and capacity-based conceptions of responsibility, on the one hand, and self-direction and -disclosure views, on the other.

Anyone who has felt the draw of Strawson's characterization of responsibility in terms of the "participant reactive attitudes," as have I, is liable to feel a bit deflated when they first come across the distinction, central to the answerability theory, between *basic responsibility* and *substantive responsibility.* Here is the pair of passages in which the distinction first appears in *What We Owe to Each Other:*

> Questions of 'moral responsibility' are most often questions about whether some action can be attributed to an agent in the way that is required in order for it to be

> a basis for moral appraisal. I will call this sense of responsibility [*basic responsibility*][8]. 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.
>
> Questions of responsibility can also be asked in other senses. For example, it might be asked whether it is a father's responsibility to set up his estate in such a way as to prevent his grown son from making foolish financial decisions. […] These judgments of responsibility express substantive claims about what people are required […] to do for each other. So I will call them judgments of *substantive responsibility.* (Scanlon 1998, pp. 248-49)

Briefly stated, then, basic responsibility is an agent's responsibility *for* something. It concerns the manner in which something, such as an action, must be "attributable" or related to the agent in order for her to be praised or blamed for it (i.e., "morally appraised" on its basis).[9] By contrast, a substantive responsibility is, in a word, a *duty.* Substantive responsibilities are the responsibilities an agent *has*—to tell the truth, to treat others with goodwill, and so on. That these two notions of responsibility come apart is most easily seen from the fact that agents can be responsible for morally neutral actions. In other words, an agent can still be responsible for something she does—such as voluntarily raising her hand—even if that thing bears no relationship to her substantive duties or obligations.[10]

The reason this distinction can be somewhat deflating for those who've taken a cue from Strawson is that it draws a very bold line between the conditions of responsibility, on the one hand, and the conditions under which an agent is legitimately subject to normative demands, on the other, whereas Strawson saw a definitive link between the two. According to Strawson, an individual's eligibility for attitudes like resentment or indignation depends on whether she is *an appropriate object* of the basic demand for goodwill. Here it is important to remember that such attitudes, in addition to being *reactive* in that they are responses to the attitudes of others, are

[8] Basic responsibility was originally "responsibility as attributability". To avoid confusion with another concept by that name, I stick to using Angela Smith's (2015) terminology. See footnote 3 above.

[9] Scanlon's choice to speak in terms of "moral appraisal" has no doubt helped contribute to the impression that the answerability theory is a theory only of what Watson (2004) calls the *areteic* face of responsibility, i.e., evaluations of moral character. But as Scanlon understands it, moral appraisal is the basis of all the ordinary responses associated with responsibility, including those associated with accountability.

[10] See, for example, Fischer & Ravizza (1998), p. 8, fn. 11.

also attitudes "of involvement or participation in a human relationship" (Strawson 1962, p. 9). That is to say, they are the attitudes we feel toward others *qua* participants in norm-governed, interpersonal relationships—individuals who, in virtue of their participation in those relationships, *have a responsibility* to conform to certain expectations.

The significance of this point is best illustrated by Strawson's remarks on what have since come to be known as *excusing* and *exempting* conditions.[11] Generally speaking, excuses are factors which call into question an agent's responsibility, praiseworthiness, or blameworthiness for a specific action or attitude, where common examples include appeals to ignorance, coercion, or the fact that the action was unintentional or that, contrary to appearances, it actually reflected a good intention.[12] Exemptions, by contrast, concern the nature of the agent. Specifically, they concern whether the agent is even "the right sort to be evaluated in terms of norms of moral praise and blame" (Vargas 2013, p. 113). Consider, for example, that when Strawson contrasts these two types of conditions, he negatively characterizes excuses as follows:

> They do not invite us to view the *agent* as one in respect of whom these attitudes are in any way inappropriate. […] They do not invite us to see the *agent* as other than a fully responsible agent. […] They do not suggest that the agent is in any way an inappropriate object of that kind of demand for goodwill or regard which is reflected in our ordinary reactive attitudes. (Strawson 1962, p. 8; emphasis in original)

Exempting conditions, by contrast, do "invite us to suspend our ordinary reactive attitudes towards the agent", and when we "see someone in such a light as this, all our reactive attitudes tend to be profoundly modified" (Strawson 1962, p. 9). That is, due to certain disabling psychological conditions[13] or particularly damaging formative experiences[14], some agents are

[11] As discussed in Vargas (2013, p. 112), Watson (1987) identifies this distinction in Strawson's work, although Strawson himself did not speak in terms of excuse and exemption.

[12] Such as Strawson's example of a person who steps on your foot to kill a venomous spider.

[13] Strawson lists being "warped or deranged or compulsive in behavior" as examples (1962, p. 10).

[14] Whether damaging formative experiences can exempt is more controversial, though it is one of the examples Strawson mentions. For a discussion of why formative experiences may only lessen the severity of moral criticism, but do not exempt, see Smith (2008), especially p. 390.

altogether exempt from the attitudes of participation and must be viewed, as it were, more "objectively."[15]

Exempting conditions thus mark the point of transition from the *participant stance*, wherein we countenance someone second-personally, as a co-participant in norm-governed, interpersonal relationships, and the *objective stance*, wherein we view someone from a technical or strategic standpoint. When we view someone objectively, our experience may be "emotionally toned"—Strawson mentions fear and pity as examples—but it may not include "the range of reactive feelings and attitudes which belong to involvement or participation with others in inter-personal human relationships" (Strawson 1962, p. 10). That is, in viewing someone objectively, we give up seeing her as a participant in norm-governed relationships, and instead we regard her as the unfortunate sufferer of conditions or circumstances which disqualify her from the usual expectations. We may see such an individual "as an object of social policy; as a subject for what, in a wide range of sense, might be called treatment; as something […] to be managed or handled or cured or trained" (Strawson 1962, p. 9). But we may not see her as someone whom it makes sense to hold accountable to the norms of collective life.

We might summarize all of this by saying that, for Strawson, part of what it means to be a responsible agent is to be capable, in at least some generic sense, of understanding and living up to the normative expectations of interpersonal relationships. Victoria McGeer has, I think, captured this aspect of Strawson's theory perfectly:

> [G]iven that our reactive attitudes are sensitive to judgements we make about whether or not someone is a fitting recipient of these attitudes, the fact that we express them effectively communicates a good deal more. It says to their recipients that we don't despair of them as moral agents; that we don't view them "objectively"—as individuals to be manipulated or managed or somehow worked around; indeed, that we hold them accountable to a standard of moral agency

[15] I am treating exemptions as marking a clear-cut distinction between responsible and nonresponsible agents, but I am sensitive to the sort of concerns discussed in McGeer (2010, fn. 2)—namely, that "the distinction between those who are fit to be held responsible and those who are not is hardly black and white" because "[r]esponsible agency involves capacities that can be more or less well-developed". These developmental considerations notwithstanding, there remains a categorical sense in which a creature either *is* or *is not* a responsible agent, depending on whether it possesses the requisite capacities to any extent. Some creatures are clearly exempt in this categorical sense.

> *because we think them capable of living up to that standard*. (McGeer 2014, p. 76; emphasis mine)

On Strawson's account, then, being a responsible agent means being legitimately subject to, because capable of fulfilling, a general normative expectation or demand to treat others with good will. It means being capable of having what Scanlon calls a *substantive responsibility*.

It would thus appear that Strawson's account runs afoul of the distinction between basic responsibility and substantive responsibility. For although it may very well be true that humans have, and so are capable of fulfilling, a substantive responsibility to treat others with good will, this is not what makes them responsible *for* what they think and do. It is, rather, a responsibility they *have.*

I believe Strawson was right to characterize (basic) responsibility as depending, in part, on the agent's capacity to understand and live up to norms, and in the next section I will explain why by considering in more detail the relationship between basic responsibility and substantive responsibility. As I'll discuss, Angela Smith, who has given what is probably the most complete presentation of the answerability theory, describes the relationship between these two senses of responsibility in counterfactual terms: what it means to say that an agent is basically responsible for something is that the agent is related to that thing in such a way that responses like praise or blame *would* be appropriate *if* that thing exceeded or violated one of her substantive responsibilities (Smith 2015, p. 106). However, there is something ambiguous about this account. Specifically, it isn't clear which "modal side" the agent's *capacity* to have responsibilities is on. Must the agent actually be the sort of agent who can have substantive responsibilities? Or must she simply be related to her conduct in such a way that, if she *were* such an agent, then she could be praiseworthy or blameworthy for what she does? In resolving this ambiguity, it will become clear why Strawson defined responsibility as partly requiring the capacity to understand and live up to the demands of interpersonal relationships: without such a capacity, though the agent may be the author of her conduct, she may never be praiseworthy or blameworthy for anything she does.

Section 2 – Eligibility, In Principle

According to the answerability theory, an agent's responsibility for what she does is both unified and rather simple: An agent is responsible for something just in case that thing reflects her judgment, or, what is the same, just in case the agent is an intelligible target of justificatory requests regarding that thing:

> [T]o say that 'A is morally responsible for X' is to say that A is an intelligible target of requests to justify X and that she is eligible, in principle, for a range of moral responses (from positive and negative appraisal, to the reactive attitudes, to attitudinal and behavioral expressions of praise and blame) based on the nature of the thing in question and the quality of the reasons she could give in response to these justificatory requests. (Smith 2015, p. 106)

The theory thus paints a fairly elegant picture of responsibility. An incredibly wide range of normative responses—from privately and coolly held judgments of moral character, to emotionally charged, public expressions of indignation and blame—are said to rest on a single common basis: the quality of the agent's judgment.

Much of that elegance is owed to the way the theory distinguishes between the following two questions: (a) the question of an agent's basic responsibility for what she does, i.e., her eligibility for normative responses like praise and blame, and (b) the question of whether any such response is actually appropriate in the given scenario, and if so, which one. This is what Smith is marking in the above passage when she speaks of agents as being "eligible, *in principle,* for a range of moral responses", about which she says this:

> My use of 'in principle' clauses here is no doubt annoying, but it is necessary. For on this view, a person's answerability for something does not depend upon whether there is in fact anyone in the circumstances who can legitimately request that she give a justification for it; it depends only on whether such a request would make sense, or be intelligible. And on this view, a person's answerability for something does not depend upon whether the thing in question in fact calls for any particular moral response from others; it depends only on whether such a response would be warranted if it should turn out that the thing in question (together with the agent's rational justification for it) exceeds or violates certain moral demands, norms, or expectations. (Smith 2015, p. 106)

That is to say, whether it is in fact appropriate, all things considered, to hold an agent responsible for something, or to hold her responsible in a specific way, or by a specific person, depends on a host of factors which "go beyond" and "have nothing to do with" the agent's basic responsibility for what she does (Smith 2015, pp. 108, 120). This is how Smith is able to resist the idea, central to pluralistic theories of responsibility, that "different moral responses may well presuppose different agency conditions" and that there are "different concepts of moral responsibility corresponding to each distinct type of moral response" (Smith 2015, pp. 104-5). In Smith's view, the sort of factors which pluralists identify as reasons to posit distinct types of responsibility are certainly relevant when considering the all-things-considered appropriateness of a particular response; but they are not relevant for determining the agent's eligibility for normative responses in general.

Notice, however, that one of the factors Smith identifies as "going beyond" basic responsibility is whether "it should turn out that the thing in question (together with the agent's rational justification for it) exceeds or violates certain moral demands, norms, or expectations" (Smith 2015, p. 106). This is important because it begins to bring into focus the relationship Smith sees between an agent's basic responsibility for what she does and the substantive responsibilities that she has. In short, that relationship is counterfactual: basic responsibility picks out the manner in which an agent must be related to her actions and attitudes, such that she *would be* praiseworthy or blameworthy for them, if they *were* to exceed or violate one of her substantive responsibilities. And it makes sense that Smith would raise this point as an objection to responsibility pluralism, since one reason that pluralists commonly provide for thinking that *accountability* is a distinct type of responsibility is that, unlike attributability, it is essentially concerned with an agent's obligations to others. Both Watson (2004, p. 231) and Shoemaker (2015, p. 113), for example, cite morally neutral actions as instances in which an agent is responsible for what she does in the attributability sense, but not in the accountability sense. Smith's claim, then, is that an agent's responsibility for morally neutral actions rests on the same basis as her responsibility for morally praiseworthy or blameworthy ones. The only difference is that, in the latter case, the action is also one which she has a substantive responsibility to perform or avoid.

There is, I think, a sense in which this is perfectly correct. An agent's blameworthiness certainly does depend on whether she has actually violated some norm that she is subject to. If it turns out that the action was, as Strawson says, "quite consistent with the agent's attitude and intentions being just what we demand they should be" (1962, p. 8)—if, put simply, she did nothing *wrong*—then the agent is not blameworthy for that action. But she may still be *eligible* for responses like praise and blame, inasmuch as blame or a similar response *would be* appropriate if the action had been inconsistent with those demands. Put in another way, if we want to know whether an agent is responsible for something in the way that is required for normative responses like praise and blame, it should not be necessary to know (a) what substantive responsibilities she actually has, if any, or (b) what the normative quality of her judgment and behavior actually is—i.e., whether she has actually exceeded or violated any of her substantive responsibilities.

But there is another way of interpreting the "in principle" qualifier which is more problematic. From what has been said so far, Smith's use of that qualifier could be understood as suggesting that, if we want to know whether an agent is responsible for something, it is not even necessary to know (c) whether the agent is *capable* of having any substantive responsibilities. In my view, this is the issue which really must be wrestled with in order to adjudicate difficult cases of psychological impairment (like the psychopath) and the disagreements they give rise to, because in such cases the contested issue is not whether the agent's behavior is harmful and in violation of ordinary ethical norms (c.f. Nelkin 2015). Rather, the issue is whether a psychopath is even *subject* to those norms, given the clear sense in which such an agent fails to meet the sort of capacity requirements that plausibly make those norms justified for the rest of us.

Now, it should be mentioned at the outset that Scanlon and Smith altogether reject this idea. That is, they do not think that an agent's moral responsibilities to others depend for their justifiability on whether that agent is capable of understanding and living up to them. This issue will take center stage in Sections Three. For now, I want to suppose for sake of argument that there could be an agent who is answerable for what she does, but is incapable of having substantive responsibilities. This would be an agent who acts in ways that reflect her judgment, but who, due to some kind of disabling condition, is not an appropriate object of normative demands. In other words, I want to beg the question for a moment against the answerability

theory's claim that answerability is sufficient for moral responsibilities, in order to clarify the conceptual connection that I see between basic responsibility, on the one hand, and the capacity to have substantive responsibilities, on the other. Then, in Section Three, I will deal directly with the question of what is required for this latter capacity, and give my reasons for thinking that answerability falls short.

So, suppose that Magdalena and Miguel both enjoy terrorizing their younger brother, Gadiel, for the simple reason that it is good fun to watch Gadiel cry. And let us further suppose that, unlike Miguel, who is psychologically ordinary in this respect, Magdalena was born with a condition which makes her incapable of having substantive responsibilities. Whatever basic psychological or agential requirements there are for substantive responsibilities, she doesn't meet them. Before saying anything about what these requirements might be, I think it can be made perfectly clear why they are relevant. Whereas both Miguel and Magdalena act in accordance with their judgment that it is good fun to watch Gadiel cry, Magdalena's judgment does not falter against any norms *which she is subject to.* The contrast between Magdalena and Miguel suggests that there are two different senses in which an agent can be said to exercise normatively faulty judgment. Magdalena's judgment is faulty in that it falls short of the ordinary interpersonal norms—i.e., the norms human agents are generally expected to conform to, but which Magdalena is exempt from in virtue of her condition. Miguel's judgment, by contrast, is faulty in the further sense that it falls short of norms which Miguel actually has a responsibility to conform to. To use McGeer's phrasing from earlier, we do not despair of Miguel as a moral agent, and so our normative evaluation of him carries some additional content. Namely, it carries the further imputation that Miguel has what it takes to live up to the norms in question and, as such, is legitimately subject to those norms in a way that Magdalena is not.

Now, even if they accepted the stipulations of the case, answerability theorists could argue that an agent like Magdalena is still eligible, in principle, to be praised or blamed for what she does, because the following is still true: if Magdalena *were* an appropriate object of normative demands, then she *would be* praiseworthy or blameworthy for exceeding or violating them. Her disabling condition may count as a reason to withhold particular blaming responses, but it does not count as a reason to think that she is not still *eligible* for such responses inasmuch as she is still the author of her own conduct.

But here is the problem. This response suggests that Magdalena is eligible for responses like praise and blame because of what would be the case, if a condition were met *which cannot possibly be met by an agent of her sort.*[16] Just as, given the sort of agents humans actually are, it is not possible for someone's natural height or aural abilities to reflect her judgment (Smith 2008), so it is not possible, given the sort of agent Magdalena is, for her behavior to reflect praiseworthy or blameworthy judgment. That is to say, while her judgment can evaluated against whatever norms we like, and be found faulty or excellent in any number of normative domains, it is not possible to evaluate her judgment as having violated or exceeded one of Magdalena's substantive responsibilities. Inasmuch as it is therefore *impossible* for Magdalena to be praiseworthy or blameworthy, it is rather difficult to maintain that she is nonetheless "eligible, in principle," for praise and blame.

So, while I think Smith is right to say that we do not need to know *which* substantive responsibilities an agent has, if any, in order to know whether she is basically responsible for what she does, I think this shows that we do need to know whether the agent is at least *generically capable* of having substantive responsibilities at all. In light of this, I suggest that the conditions of basic responsibility be expanded to include the following capacity requirement: the agent must be related to her own activity in the sort of way that makes her generically capable of having substantive responsibilities—so that she could, in principle, *have a responsibility* to conform to the prescriptions of norms.

Before I go on to say what I think this capacity requires, and my reasons for thinking it requires more than answerability, I want to emphasize that my principal aim in this section has been to argue that the capacity to have substantive responsibilities (whatever this should turn out to be) is a necessary condition for basic responsibility. Even if rational self-direction (answerability) were sufficient in this respect, it would still be true that part of what explains why an agent is basically responsible for what she does is that she has this generic capacity.

Indeed, this is how I think we should understand Strawson's claim—that, in order to be responsible for what she does, the agent must be an appropriate object of the basic demand for goodwill. Namely, we should understand it as a claim about a generic capacity to be subject to normative demands at all, as opposed to a claim about any particular demand. And I think it is

[16] Making it a *counterpossible* condition. See Berto & Jago (2018), Section 1, for a discussion.

quite plausible that Strawson meant for the "basic demand" to be understood in this generic sense, seeing as he was careful to choose a purely formal characterization of it: as a demand to treat others with goodwill. What, exactly, that goodwill amounts to will depend on the concrete details of the relationship. He says, "In general, we demand some degree of goodwill or regard on the part of those who stand in these relationships to us, though the forms we require it to take vary widely in different connections" (Strawon 1962, p. 7). But if all interpersonal relationships are, as seems plausible to say, governed by *some* normative expectation, then all of them will feature some form of the basic demand for goodwill. And so any creature (including potentially a human being) which lacks the generic capacity to be subject to norms would be exempt from the whole range of responsibility-relevant responses, just as Strawson suggests. Importantly, this would be true *even if* there remains an objective or technical sense in which such a creature's judgment (or will) has been normatively flawed. As Pamela Hieronymi has recently put this idea: it is not that such a creature doesn't really mean us harm, or know what it is doing, it is that "the ill will does not matter in the usual way" (Hieronymi 2020, p. 10).

Section 3 – The Intelligence of Responsibility

Very much in opposition to this strand of Strawson's theory, answerability theorists maintain that any agent who can act for her own reasons is subject to moral norms, even if she is incapable of understanding and living up to them. For instance, Smith and Scanlon have both argued that agents like psychopaths and "hardened" criminals are still morally blameworthy, as long as their behavior reflects their own morally faulty judgment. In Scanlon's words:

> Whenever a person's self-governance is faulty […], a judgment that this is so is correct, which is to say that he or she is subject to justified moral criticism [blame].
>
> […]
>
> [A] rational creature who fails to see the force of moral reasons […] can nonetheless understand that a given action will injure others and can judge [incorrectly] that this constitutes no reason against so acting. […] A person who is unable to see why the fact that his action would injure me should count against it still holds that this *doesn't* count against it. (Scanlon 1998, pp. 269, 288)

On the answerability theory, then, even a "morally blind" agent has a responsibility to conform to moral norms and may therefore be normatively appraised (praised or blamed) for how well

she conforms to them. Further articulating this idea, Smith has said that the difference between such an agent and a vicious dog is that "we do not think that a dog's disposition to attack us reflects a judgment that we lack moral value or standing, or indeed that it reflects any judgment at all on the dog's part about what reasons it has" (Smith 2008, p. 388). Whether the agent is also capable of seeing that her judgment is mistaken and modifying it accordingly is, on this view, entirely beside the point, because our relationship with such a person is already impaired by the normatively faulty attitude she has toward us (Scanlon 2013, pp.86-92).

Answerability theorists thus reject the possibility that an agent could exercise morally faulty judgment without also being *at fault* for this failure. I think this is a mistake. In what follows, I approach this issue somewhat differently by considering some general features of capacities, and the different ways agents are open to being evaluated for how well they exercise their capacities. My goal is to bring some precision to the idea that was suggested earlier by the contrast between Magdalena and Miguel—namely, that there could be an agent who is answerable for what she does, and can even be normatively evaluated from an "objective" standpoint, but who is nonetheless an inappropriate object of the normative demands that structure ordinary, interpersonal relationships.

I want to begin with the simple observation that judgments, like all activities of agents, are exercises of a capacity—in this case, a capacity to draw practical conclusions, to recognize and respond to reasons, to govern oneself according to principles, and so on. To evaluate an agent's judgment is therefore to claim that the agent has exercised her capacity for judgment well or poorly by certain normative standards, i.e., in a way that it should or should not be exercised.

Capacities, in the broadest possible sense, are the dispositions of a thing to respond to inputs, stimuli, or circumstances in various ways. For any disposition, there are certain standards of responsiveness, usually indicated by the truth of certain counterfactual claims, which something must conform to in order to be correctly attributed with that disposition. Even inanimate things—sticks, stones, coffee mugs, and clothes lines—have dispositions. They have dispositions to break when bent, to sink in water, to shatter when dropped, and to sag when weighted. To say that standards are at issue even with respect to the attribution of these mundane dispositions is to say, for one, that we would not describe something as, say, *buoyant* if it under no circumstances floated in a liquid or gaseous medium. And it is likewise to say that things

possessed of these dispositions can be graded in terms of how well or poorly they conform to the ideal, just as we can speak of a substance as being an *excellent* catalyst or a *poor* conductor of electricity. The word "capacity," however, does tend to evoke connotations of agency in a way that these mere dispositions, as we might call them, do not. Capacities tend to be understood as dispositions to *do* something.[17] These specifically agential capacities are the ones I would like to discuss.

Among agency-involving dispositions there is yet another important distinction. In *The Concept of Mind* (1949) Gilbert Ryle argued that certain agential capacities are merely routinized or habitual, whereas others are "intelligent." An intelligent capacity, in short, is a capacity not only to conform to certain dispositional standards, but to *apply* those standards, and not only to apply them, but to *adjust* the standards one applies in response to corrective feedback. This would be a capacity, for example, not only to play a game in a manner consistent with its basic rules, but also to learn to play it better, adopting new strategies and revising old ones in light of experience and instruction. Unlike habits, intelligent capacities are not just dispositions to regularly think and act in certain ways. They are "higher-grade dispositions" to regularly *update* one's thought and behavior, to display a degree of diachronic flexibility and a sensitivity to specific cues. Victoria McGeer (2018) summarizes their three central features as follows:

> [Ryle's] first (and most often repeated) observation is that these dispositions are invariably 'multi-track', or as Ryle himself puts it, these are dispositions 'the actualisations of which can take a wide and perhaps unlimited variety of shapes' (1949, p. 44). His second observation is that in exercising such capacities, people are responsible for what they do in a distinctive kind of way; for example, they show themselves 'ready to detect and correct lapses, to repeat and improve upon successes, to profit from examples of others and so forth' (Ryle 1949, p. 42). In short, they work at honing and developing such capacities even as they exercise them, thereby instantiating a dynamic form of 'reasons-responsiveness'. His third observation is that feedback from the environment is an essential component of this dynamic process, and preferably feedback of a relatively distinctive kind—what Ryle calls 'training', as against mere 'drilling' or conditioning (1949, pp. 42–3). For not only must the agent get the message that she is performing well or badly, she must get this message in a way that encourages her to think critically about how to improve upon her own performance, thereby developing the 'intelligence' of her intelligent capacity. (McGeer 2018, p. 351)

[17] Whereas the processes of change undergone by coffee mugs and clothes lines do not qualify as *doings.* They are mere *happenings* (Nagel 1986).

These three, interrelated features may be itemized as follows:

1. An intelligent capacity is *multi-track* in the sense that it can be exercised in a wide variety of ways, even in response to the same sort of inputs.

2. An intelligent capacity is dynamic in the sense that it tends to change, and not merely change but *improve*, as it is exercised.

3. An intelligent capacity tends to improve because it is sensitive to *corrective feedback* from the environment, including especially the social environment.

As the second and third of these features make clear, intelligent capacities bear a unique relationship to their dispositional standards. First, to say that they tend to improve is, of course, to say that there is some standard relative to which an exercise of that capacity can be evaluated as *faulty*, *satisfactory*, or *excellent*, and, more importantly, that the capacity tends to perform better by those standards with time. Second, to say that intelligent capacities change in response to corrective feedback is to say that the agents who have them are sensitive to features of their environment, including the consequences of their own performances, the examples set by others, and the instructions and evaluations they receive from peers and mentors—all of which serves to cue the agent into possible areas of improvement.

We can get an intuitive idea of what the intelligence of intelligent capacities consists in by reflecting on the familiar phenomenon of *learning*. Consider the following contrast. A person who today makes a rookie mistake during a game of chess may nonetheless, with the requisite practice and training, grow into a grand master. At the very least, it is likely that just this one blunder will be enough to cue her into the fact that some aspect of her playstyle may be in need of improvement, especially if the nature of her mistake is made clear to her through example and instruction. On the other hand, a simple AI which has been programmed to play chess by way of an especially inflexible algorithm—one which is set up to perform exactly one move in response to each possible configuration of the chess board—can also make mistakes. It can make mistakes in the sense that it makes suboptimal moves, moves which, say, a grand master or a more sophisticated AI would avoid.[18] Crucially, however, this simple AI is incapable of learning from

[18] It could also make mistakes in the sense that it fails to correctly execute its chess program—fails to *perform* up to its *competence* (Chomsky 1965). For simplicity, I am assuming that the

its mistakes. It cannot appreciate that its moves have sometimes been suboptimal. Even now that it has lost the match, if it were presented with the same series of board configurations, it would respond in exactly the same manner and lose the match again.

Of course, a software engineer can reprogram the AI so as to correct these mistakes. This resembles formal instruction somewhat, but it is clearly distinct from the form of learning which the novice human player is capable of. The player *herself* is poised to register her mistakes as mistakes, to be cued into these mistakes by their consequences and by what others tell her, and, finally, to adapt her playstyle in ways she believes will avoid similar blunders in the future. The kind of control she has over her chess playing does not consist *merely* in the fact that she is the one making the moves—i.e., on the basis of her own assessment of the situation. The AI possesses this sort of control as well. The nature of her control also consists in the fact that her playstyle is something *she* can change in light of how it has "played out."

The difference between the AI's capacity to play chess and the novice player's can be made more precise by considering how each would look if modeled as a *response function*, with configurations of the chess board as the inputs and chess-related judgments and actions as the outputs (see Figure 1 below).

Figure 1. The figure above depicts a simple block diagram representing the relationship between inputs and outputs for the chess AI.

---

chess AI is idealized in three respects: first, that it never fails to correctly execute its program; second, that it never wears down or otherwise falls into disrepair; and third, that its programming does not feature any random values (it never flips a coin to decide what to do). It is a "tropistic creature" *par excellence* (Dennett 1981, Chapter 5).

The AI's response function is *static*, in that the mapping of board configurations, on the one hand, to judgments and actions, on the other, is diachronically invariant. Save for direct intervention by a software engineer into the AI's programming, the AI will respond to each board configuration in the same manner every time it is presented with that configuration. This is why, if the AI loses a match and were then presented with the same series of moves, it would lose the match again in exactly the same way.

The novice player's response function, by contrast, is *dynamic*. She does not always respond to the same board configuration in precisely the same way, because her capacity to play chess has certain *second-order dynamical properties*. The second-order dynamical properties of a thing refer to its disposition to change its response function over time, or, "how its dispositions to respond to various stimuli varies from one moment to the next" (Ismael 2016, p. 27). Unlike the AI, the human player is sensitive to environmental feedback that calls on her to adjust how she interprets and responds to the chess board. Where she previously thought it was obviously best to take her opponent's pawn in one sort of circumstance, now she recognizes that this response will have undesirable consequences, due to factors which she didn't previously recognize or respond to—such as the position of her opponent's knight, which is poised to capture her assailing piece as soon as she lifts her fingers. The same configuration which she previously interpreted as an *opportunity* she now interprets as a *trap*. This is key, because it enables the novice player to learn not to be duped by the same chess moves over and over. As the saying goes, fool her once, shame on you, fool her twice, shame on her. The same simply cannot be said for the chess AI.

Now, I want to consider what it means to say that the chess AI has exercised faulty judgment with respect to certain chess-playing norms, as compared to what it might mean to say that the novice chess player has exercised faulty judgment with respect to those norms. The assumption will be that there are certain standards, or norms, which govern which situational features one should recognize as chess-related reasons (norms of recognition or detection) and

how one should decide and be motivated to act on their basis (norms of reactivity).[19] A given agent's response function, then, understood as a description of her disposition to recognize and react to reasons, can be evaluated in terms of how well it conforms to these norms.

In the case of the AI, all that we can possibly mean when we say that it has exercised faulty judgment with respect to these chess-playing norms is that its judgment is *deficient.* That is to say, its response function has not been in conformity with the norms of chess, but in a way that carries no further imputation of *fault* or *culpability*. This is the sort of evaluation the AI's software engineer is apt to make as she considers the ways in which her creation might be improved. "Ideally," she might say to herself, "the AI would recognize in the present configuration of the board sufficient reason to make a particular move, M, whereas actually it recognized in that configuration sufficient reason to make another move, M*." In light of this observation, the software engineer can say, quite correctly, that the mapping from board configurations to moves is not as it should be—that, in a purely technical sense of the phrase, the AI "should have done otherwise."

Of course, the novice chess player can be said to exercise faulty judgment in this same sense.[20] Insofar as her response function is not in conformity with the norms of chess, we can correctly observe that she has exercised her capacity to play chess deficiently. She "should have done otherwise." But in voicing this sentence we might also—and probably do—mean to communicate a good deal more. We might also mean to say that the normative ideal is something *which she herself is capable of*. Not only is her response function not in conformity with norms of chess playing, but it is also not in conformity with her own abilities. Thanks to the second-order dynamics of her capacity to play chess—i.e., because hers is an intelligent capacity, rather than a routinized one—she herself has the power to conform to normative standards of which she presently falls short. This mirrors McGeer's point from earlier: that the reactive attitudes communicate to their recipients "that we hold them accountable to a standard of moral agency because we think them capable of living up to that standard" (McGeer 2014, p. 76). Here the

[19] Here I follow Fischer and Ravizza's distinction between an agent's receptivity, or ability to recognize reasons, and an agent's reactivity, or her ability to be appropriately affected by reasons (Fischer & Ravizza 1998, p. 41).

[20] It is a "resource" we can make use of, as Strawson says (1962, p. 10).

standards are different—they are standards of "chess agency," so to speak—but the idea is the same.

To say that the novice player "should have done otherwise" in this more robust sense is to make a claim about her that cannot correctly be made about the chess AI. When the novice player's peer or mentor says (to her) that she "should have done otherwise", the "ought" in this sentence implies the "can" of *capacity*—something the novice player is capable of doing. The AI's failure, by contrast, can only be chalked up to its bad programming. When the software engineer says that the AI "should have done otherwise", the "ought" implies the "can" of *possibility*—something that could be different about the AI, a change it could (passively) undergo or a way it might have been different. This kind of talk is ubiquitous in the realm of artifacts. Your car shouldn't have continued accelerating when the gas pedal was release. The lights should have come on when the switch was flipped. It is a perspective we can take even on our bodies, as when a physician remarks that her patient's lungs are not functioning as they should, or on entire organisms, as when a zoologist observes that an ant is not acting as it should due to a parasite. Perhaps there is not Aristotelian teleology in the natural world, but there is at least enough functionality to speak sensibly of the ways things should be and should behave.

There are thus two very different ways in which agents can be normatively evaluated for how well they exercise a capacity, each with its own conditions of eligibility. In order to be evaluated from a technical standpoint, the agent need only possesses the first-order capacity in question. The chess AI is disposed to regularly respond to configurations of a chess board in a manner that suffices for the ascription of a capacity to play chess (whereas, by contrast, we would not ascribe that capacity to the AI if it responded to different configurations of the board with different lines from Shakespeare, as this is not how chess is played). Insofar as the AI has in fact exercised this capacity, then, it is open to being technically evaluated for how well it conforms to the norms of recognition and reactivity associated with that capacity.

By contrast, what makes an agent eligible for the mode of normative appraisal that is at issue in responses like praise and blame—responses, namely, that evaluate the agent as having exceeded or violated a norm which she has a responsibility to conform to—is that she *has what it takes* to understand and live up to those norms. As we have seen, the novice player has more than a first-order capacity to recognize and respond to chess-related reasons; she also possesses a

second-order capacity to bring her responsiveness up to higher standards than it presently conforms to. She has what might be called *dynamic control* over how well she recognizes and responds to her situation, which exceeds the static form of rational self-direction that the AI is capable of. Whether she is *in fact* under any obligation to play better chess is an altogether separate question, just as Smith suggested in the passage from Section Two. But here the crucial difference is that, in addition to acting in ways that reflect her judgment, part of what makes the novice player eligible (in principle) for praise and blame is the fact that she is sort of agent who *could be* under such an obligation, whereas the routinized chess AI is not.

Section 4 – Making Things More Familiar

In the previous section I drew a distinction between two ways in which agents can be evaluated for how they exercise their capacities, only one of which is a plausible candidate for the manner of evaluation that is at issue in our praising and blaming practices. But so far I have only focused on the capacity to play chess and make chess-related judgments, and one of the agents in my case is only artificially intelligent. Still, I think that the distinction between these two modes of normative evaluation remains in force even as we expand the case to include judgment in general, and even as we add certain "humanizing" features, such as the phenomenal experience of what it is like to make these judgments. Scanlon has even suggested something along these lines in his response to an objection that his notion of rational self-governance is too weak:

> [A] sophisticated computer that was programmed to weigh evidence and balance competing reasons might be said to 'govern' its outputs […]. Such a machine would be 'responsible' in a causal sense for the processes it governs. We would say that errors in its program are 'responsible for' defects in its output. But we would not regard it as 'responsible' in the sense required for moral blame.
>
> This objection draws its plausibility from two presuppositions. The first is that there would be no point in expressing moral indignation or blame to a computer (even a very sophisticated one of the kind imagined) or in engaging in moral argument with it. This would be like pleading with your alarm clock. But this inappropriateness derives not from the fact that a computer is a causal mechanism but *rather from what are assumed to be the limited forms of interaction that we can have with it*. If […] we are 'like computers' in the very general sense that our mental lives depend on underlying causal processes, *it will nonetheless remain true that we can communicate with each other in moral terms and that our*

> *behavior will be influenced by this kind of dialogue in just the way it is now*. (Scanlon 1998, pp. 281-82; my emphasis)

I think that this is exactly right. Unfortunately, in the final analysis, Scanlon is not committed to thinking that the capacity to learn from interpersonal exchange and other forms of corrective feedback is a condition of responsibility. But in my view, the above passage is right to suggest that the required forms of interaction go well beyond requests for the agent's reasons (which a much less sophisticated computer would be well equipped to fulfill). The point at which we would be justified in feeling blame toward a computer is precisely the point at which the computer begins to behave like a dynamic agent—an agent who isn't just running a script, but who has read and write privileges on at least some of the script it is running (Millgram 2014). This kind of responsiveness is what opens the door to normative expectations and, where they are not met, normative criticism and injunctions to do better. It is what opens the door, in short, to norm-governed relationships. For even if a sophisticated AI could give its reasons for what it does upon request, if its capacity for judgment is static in the way I've described, we could never intelligibly recognize it as an appropriate object of demands, i.e., as an agent who *could have a responsibility* to conform its judgments to interpersonal norms. This would be an agent who, "though you may talk to him, even negotiate with him," cannot genuinely be reasoned with (Strawson 1962, p. 10).

In essence, then, I think Scanlon's and Smith's mistake, and the mistake of self-direction and self-disclosure theories more broadly, lies in thinking that the kind of agency which is sufficient for self-direction or self-disclosure is also sufficient for participation in norm-governed relationships. The harms caused by a psychopath, or by an otherwise normatively disabled agent, certainly *would* constitute an impairment of our moral relationship with her if she *were* a participant in that sort of relationship. But since she is incapable of understanding and living up to the norms which govern that relationship, and is therefore incapable of participating in it, there is no existing relationship for her to impair. Interestingly, Scanlon even floats this possibility when he says that, "Insofar as they [psychopaths] lack the capacity to understand and respond to moral requirements, it is questionable whether they can be participants in the moral relationship. […] If there is ambivalence in the case of psychopaths […] it is about whether they are actually candidates for moral relations at all" (Scanlon 2013, p. 95). This paper has, in large part, been an attempt to resolve that ambivalence.

Conclusion

This paper has tried to accomplish two aims. The first was to argue that the capacity to have substantive responsibilities is a necessary condition of basic responsibility. The second aim was to argue that the capacity to have responsibilities outstrips answerability and requires the dynamic form of reasons-responsiveness exhibited by what Ryle (1949) calls an "intelligent capacity." Any account that identifies responsible agency with a first-order capacity for self-direction or self-disclosure will fall short in the same respect. From the simple fact that someone has acted in a way that reflects her judgment, or expresses her fundamental desires, nothing at all follows about whether it would be reasonable to expect that agent to bring her thought and behavior into conformity with interpersonal norms.

That said, and to close with a gesture toward reconciliation, if the broader self-direction and self-disclosure approach were to embrace an account of selfhood as a dynamic entity with the kind of second-order properties here discussed, *this* would be a kind of agential sourcehood which could actually bear the weight of substantive responsibilities. The solution, then, is not to abandon the source-theoretic approach, or the idea that the self is centrally implicated in responsibility. It is rather to rethink the ways in which agents can be the source of their behavior and the role that selves play in the regulation of that behavior over time.

TEACHING AN OLD DOG NEW TRICKS: INTUITION, REASON, AND RESPONSIBILITY

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. 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.

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 constitute 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 2) and the challenge presented by dual-process theories (Section 3), 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 4), 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 5). 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. 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.[21] And it is capable of learning precisely by *holding each other accountable* for our reasons-responsiveness failures.

Section 1 – 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 was capable of recognizing and responding to the reasons in her situation. We can think of reasons here as any consideration which speaks for or against something, such as an action or a judgment. And we can think of responsiveness to reasons as the agent's capacity to detect and react to the reasons in her situation. So, an agent is blameworthy just in case (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, in order to justifiably praise or blame someone for something, that thing must have been "up to her" in some sense. Traditionally this was understood in terms of what the agent would have done, had something been different, such as her desires. That is, the agent could have done something else

[21] 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.

if she had wanted to. 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. The basic idea being: even if the agent could not have actually done otherwise, *she* was still the one "guiding" her behavior, and she was still a *rationally competent* agent. She did what she did as a result of her own, normal process of deliberation—just as she would have, had no such counterfactual intervener been present. In their view, this "guidance control" is all the control required for responsibility.

There are three ideas motivating the reasons-responsiveness approach, each expressing slightly different intuitions. One is that, in order to be responsible for something, surely that thing must be *attributable* to the agent, and one natural way of understanding this attributability is in terms of causal sourcehood. 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.

The second motivating idea is that, in order to be blamed for her failures, the agent must have been capable of doing the right thing. 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. If ought implies can, then Mariem must also have been capable of doing otherwise, and, more to the point, she must have been capable of doing *what she should have done.*

The third idea is closely related to the second. Keeping with the example: 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. So it is not enough that Mariem

was capable of yielding by way of any old process. She must also have been capable of yielding as a result of a "reasons-sensitive" process (McKenna 2013).

With this exposition in the foreground, it is easy to see why the reasons-responsiveness approach is challenged by the sort of dual-process model of human cognition. If most human judgment is caused by nonconscious, intuitive processes, rather than explicit deliberation about one's reasons, and if explicit deliberation mainly serves to provide post hoc rationalization of one's intuitions (confabulation), it is very tempting to conclude that most of our judgments (and the actions we perform on their basis) are not genuine responses to reasons. Human beings may *think* they do what they do for reasons, but as it turns out, 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 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.

Section 2 – 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 reigns 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 biased, nonconscious processes, then we may not be responsible for much of anything we do. In what follows I 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. 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.

Section 3 – Intuitions as Reasons-Responses

When Fischer and Ravizza originally described the capacity to recognize and respond to reasons, they used 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 could be revised 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 accessible to, conscious awareness. Here reasons-responsiveness would be understood as the agent's capacity to perceive practically salient features of her 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, although they do tend to be described as systematically biased and unreliable (Kahneman 2003, 2011; Greene 2007, 2013). Contrary to this trend, Peter Railton has argued that affective processes can be responsive to reasons, due to the manner in which some of them *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, these representations are *first-order* representations (Railton 2017, 176). The organism need not consciously represent its first-order, 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 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).[22] 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.

[22] 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.

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.

Section 4 – 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, 25-6)

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).

Section 5 – 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 2 below).

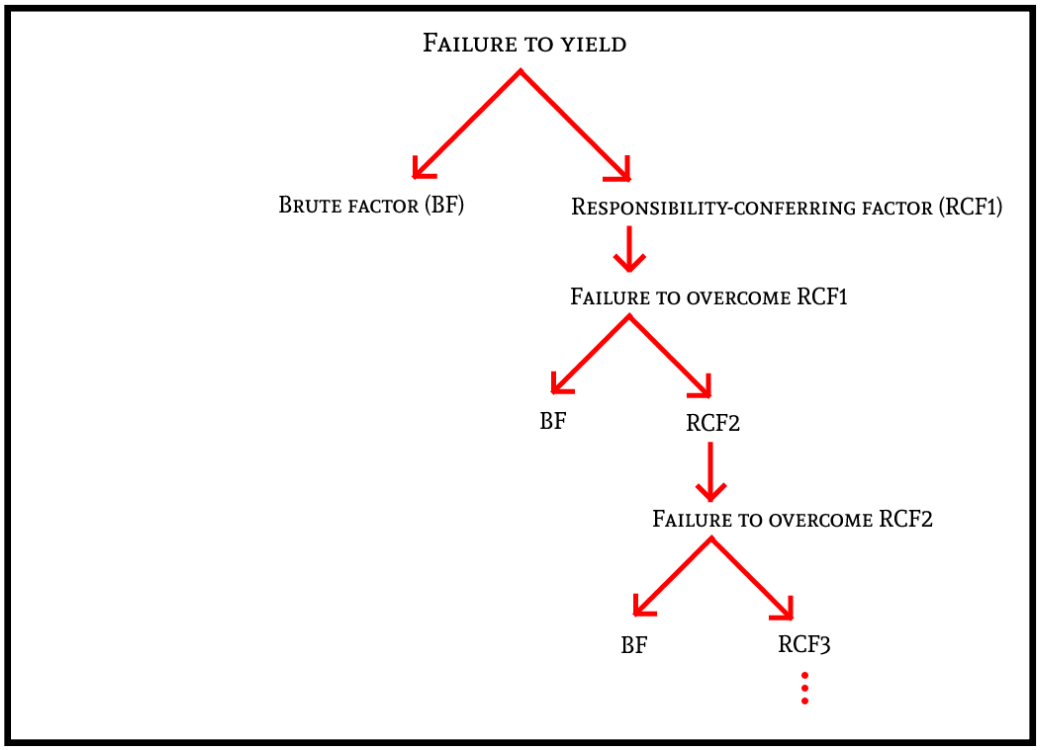

Figure 2. The figure above depicts the explanatory regress which occurs for traditional reasons-responsiveness accounts.

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*.[23] Now we will 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—and it can't 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 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. 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

[23] 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.

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.[24] 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)

But even if we accept the revised, two-tiered 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*:

[24] 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.

> 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.

In essence, my objection to Washington, Holroyd, and Kelly's ecological approach 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 constitute *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 is 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 precisely because holding them responsible partly constitutes their capacity to respond to reasons. Their standing sensitivity may even be relatively low—too low, even, to constitute traditional reasons-responsiveness; 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.

Conclusion

To bring things to a close, I want to direct our attention back to Jonathan 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.

DEEP SELVES ARE JUST SPECIAL KINDS OF REASONS-RESPONDERS

Abstract: In this paper I argue that deep self theories are just variations on a (broadly) reasons-responsiveness-theoretic framework. That the two theories bear a close relationship to one another has already been persuasively articulated by Chandra Sripada (2015), but Sripada (2017) maintains that they are different in one important respect: that reasons-responsiveness theories, but not deep self theories, consider *flexible control* to be a necessary condition of moral responsibility. Using resources from McKenna (2013), I argue that Sripada has misunderstood the function that flexibility plays in reasons-responsiveness theories, which leads him to disagree with John Fischer and Mark Ravizza's account for the wrong reasons. In my view, Sripada's real point of disagreement is with their notion of a *mechanism of action*. That is, the reason Fischer and Ravizza's account cannot explain the intuition that an agent like Harry Frankfurt's (1971) willing addict is responsible for what he does is that they identify the wrong agential features, and not because they are committed to thinking that flexibility of response is a necessary condition of moral responsibility. I conclude that the point of substantial disagreement between deep self theories and reasons-responsiveness theories is about which agential features are relevant for moral responsibility.

Introduction

In this paper I argue that deep self theories are just variations on a (broadly) reasons-responsiveness-theoretic framework. That the two theories bear a close relationship to one another has already been persuasively articulated by Chandra Sripada (2015), but Sripada maintains that they are different in one important respect: that reasons-responsiveness theories, but not deep self theories, consider *flexible control* to be a necessary condition of moral responsibility. Using resources from McKenna (2013), I argue that Sripada has misunderstood the function that flexibility plays in reasons-responsiveness theories, which leads him to disagree with John Fischer and Mark Ravizza's (1998) account for the wrong reasons. In my view, Sripada's real point of disagreement is with Fischer and Ravizza's notion of a *mechanism of action*. In fact, I believe that the point of substantial disagreement between specific deep self theories, such as Sripada's self-expression view, and specific reasons-responsiveness theories, such as Fischer and Ravizza's, is in *which agential features* are considered relevant for moral responsibility.

Section One and Section Two feature a summary of the two theories under consideration. In Section Three I summarize Chandra Sripada's reasons for thinking that deep self and reasons-responsiveness theories are difficult to disentangle. Sripada maintains, however, that an important difference remains between the two approaches. Then, in Section Four, I give the cases which Sripada believes illustrate this difference and which he uses to argue in favor of the deep self approach. In Section Five I object to Sripada's conclusion on the grounds that he has misunderstood what I call the "flexibility condition" that is part of the reasons-responsiveness approach, and that a proper understanding of that condition will reveal that Sripada's issue is really with Fischer's notion of a mechanism of action. From there I conclude that the substantial point of disagreement between deep self and reasons-responsiveness theorists should really be over which agential features are relevant for moral responsibility—i.e., which desires, psychological mechanisms, and so on.

Section 1 – Reasons-Responsiveness Theories

On the reasons-responsiveness approach, morally responsible agency requires responsiveness to reasons. We can think of reasons here as any considerations (usually features of a situation) which speak for or against some thing (usually an action).[25] A person is morally responsible[26] for what she does, then, only if she exercised, and so possesses, a capacity to recognize and react to reasons.

But how do we determine whether a person has exercised such a capacity? In order to make this determination, reasons-responsiveness theorists conduct a *modal analysis* of the person under consideration. This analysis is aimed at determining how the same person would have responded in a range of possible worlds, and (for present purposes[27]) it comprises three variables:

[25] Although the literature has traditionally focused on voluntary reactions and on deliberate actions in particular, I believe the broad contours of the reasons-responsiveness approach leave room for nonconscious reactions, including habitual actions, slips of the mind, and the formation and modification of attitudes. For simplicity's sake, however, I will speak only about deliberate actions in this paper.

[26] I will be understanding responsibility in the broadly Strawsonian sense of being an eligible or appropriate target of normative responses like the reactive attitudes and praise and blame.

[27] Two more variables are (4) the kind of relevant possible worlds (i.e., how "nearby" or similar they should be) and (5) the number or percentage of those worlds in which certain counterfactuals need to be true of the person.

1. The features which are held fixed across possible worlds
2. The features which are bracketed from our considerations
3. The counterfactuals which must be true if the person[28] is responsive to reasons

Consider Fischer and Ravizza's *actual-sequence account* of moral responsibility (Fischer & Ravizza 1998). The account is an actual-sequence account in the following sense: the features which will be held fixed in the modal analysis are those which were *part of the actual causal process* which led to the action under consideration. That causal process, which we can also think of as the way in which the action was performed, is called the *actual-sequence mechanism of action.* Correlatively, the features which will be bracketed from the modal analysis are those which were not part of the causal process leading to the action, including any counterfactual interveners or devices.

Suppose, for instance, that I have ensured that you will carry through with some intention—to lie under oath, let us say. I am implicated in the investigation, too, and so I have a vested interested in ensuring that you don't spill the beans. So I call up my neuroscientist friend and ask him to secretly install a device in your brain, which he is more than happy to do. This device is set to trigger you to decide to lie and then carry through with that decision, should you waver in your intention. As it turns out, you never waver in your intention, and so the device plays no causal role in producing your action. You lie just as you would have, had no such device ever been installed.[29]

This brings us to the final variable: the counterfactuals which must be true if a person is responsive to reasons. There are two such conditions. As we will see, only one of these conditions is at issue in Sripada's objection to the reasons-responsiveness theory, but by considering both we will gain a fuller appreciation of the basic idea motivating Fischer and

[28] Strictly speaking, it is the agent's mechanism of action which must be responsive to reasons, but for ease of exposition I will sometimes speak of the agent as being responsive to reasons.

[29] The impetus for Fischer and Ravizza's actual-sequence account was to deal with just such cases, known as Frankfurt-style cases. Although there is some controversy as to whether they were conclusive in this respect, such cases were advanced as counterexamples to the Principle of Alternative Possibilities (PAP). Intuitively, the persons in such cases are still morally responsible for what they do, despite the fact that they could not have done otherwise, due to the presence of the counterfactually intervening factor. According to Fischer, the lesson to be drawn from such cases was that moral responsibility depends only on those features of a person which actually contributed to the performance of some action.

Ravizza's account: namely, that what matters for moral responsibility is "what the agents actually do, and how their actions come to be performed" (Fischer & Ravizza 1998, p. 36), i.e., whether the process leading to the agent's action was genuinely responsive to the reasons at hand.

According to Fischer and Ravizza, it is not sufficient for moral responsibility that the person's actual-sequence mechanism would have issued in the same action if no counterfactual intervener had been present. It is also necessary that the actual-sequence mechanism would have issued in a different action in response to sufficiently different reasons. Call this the *flexibility condition.* According to this condition, an agent is not morally responsible for what she does if she "would still behave in the same way, no matter what the relevant reasons were" (Fischer & Ravizza 1998, p. 36), i.e., even if there were sufficient reason to do otherwise. If, holding fixed the agent's actual-sequence mechanism, there is no circumstance in which the agent recognizes and responds to sufficient reasons to do otherwise, then her actual-sequence mechanism is not responsive to reasons, and so she is not morally responsible for what she does in the actual case.[30] The rationale for the flexibility condition is, roughly, that agents are not morally responsible for actions they are "hard-wired" to perform, in the sense that they would perform that action no matter what reasons they were presented with. If in our above case you would have lied under oath no matter what compelling reasons there were to tell the truth—even if, say, you knew lying would lead to the gruesome death of your one true love, whereas telling the truth would have a comparatively minor consequence or none at all—then it would seem that your choice to lie would not really be chosen in response to the reasons at hand. Perhaps the behavior was the result of a compulsion, pathology, or underlying neurobiological condition, but it was not the result of a reasons-sensitive process.

In addition to the flexibility condition, it is also necessary on Fischer and Ravizza's account that the actual-sequence mechanism exhibit some degree of regularity in its pattern of response to reasons—that, when presented with reasons that are of a similar nature and similar strength, the mechanism issues in a similar action. Call this the *regularity condition.* To illustrate this condition, Fischer and Ravizza present the example of an agent, Brown, who frequently

[30] Fischer and Ravizza also discuss the range of possible worlds at which they believe this counterfactual must be true (1998, pp. 41-46), but for present purposes this aspect of the theory can be set aside. Sripada argues that it is not even necessary that the agent's actual-sequence mechanism would issue in another action at *any* other possible world—i.e., that the flexibility condition need not be met to any extent—and my goal is to show why this is not true.

chooses to take a drug called "Plezu." Brown would not take Plezu, however, if it cost $1,000 per injection. He thus appears to be responsive to at least some reason to do otherwise. However, Brown's responsiveness to reasons is irregular in that he "would *only* recognize the thousand dollar price to be a sufficient deterrent" (Fischer & Ravizza 1998, p. 70). That is, if the price of Plezu were, say, three or four thousand dollars, or even $1,001, he would take the drug. As such, although Brown does otherwise in at least one circumstance, it would appear that process leading to his action is not really *connected to the reasons* in that circumstance. For if Brown had refused the drug *because* of its exorbitant price—i.e., if his actual-sequence mechanism had actually been responsive to this reason—then he should likewise refuse the drug when the price is even greater.[31] The fact that he does not suggests that Brown's flexibility of response is due, instead, to some kind of fluke, glitch, or pathology.[32] This is why Fischer and Ravizza maintain that the agent must exhibit an "understandable pattern" of response—one that is "grounded in reality" (Fischer & Ravizza 1998, p. 71, 73).

Taken together, the truth of these two counterfactual conditions is meant to ensure that the person's action really is a consequence of (a) there actually being some reason to perform that action and (b) her recognition of and reaction to this reason. That is, they are meant to ensure that her action is not merely the consequence of something like a compulsion, a neural misfiring, or a mental disorder that is breaking the connection between reasons and her action. What makes a person morally responsible for some action, then, is that she performed it as a result of her own reasons-sensitive process. She must *be the one guiding the action*, and she must be guiding the action in a way that is *actually connected to the reasons* in her situation.

Section 2 – Deep Self Theories

According to deep (or real) self theories, by contrast, a person is morally responsible for an action just in case that action is expressive of her deep self. The intuitive idea here is that a

[31] This is not to say that Brown's selective response to the $1,000 price could not be part of an understandable pattern. For example, if there was something about that price in particular, but no price higher or lower than it, which counted against using Plezu—perhaps the $1,000 price is actually a signal communicating the presence of law enforcement—then Brown's response would have been due to a reasons-sensitive mechanism after all. Fischer and Ravizza flag this possibility in their discussion of the "saber killer" example (1998, pp. 65, 72-3).

[32] Fischer and Ravizza speak of such irregularity of response as "evidence of [the agent's] insanity" (1998, p. 65).

person should only be considered morally responsible for things which reflect who she really is, things with which she can in some sense be identified. How the self-expression relation gets defined and what the deep self amounts to varies between theorists, but common to all deep self theories is the idea that there is a special subset of a person's psychology with which she can "really" be identified. Often that subset is a subset of the person's *desires*. These could be, for example, desires resulting from coolly deliberated all-things-considered judgments (Watson 1975), second-order desires about which first-order desires one wants to act on (Frankfurt 1971), or they could be one's fundamental, abiding desires—or, *cares* (Sripada 2015, 2016). In order to morally responsible for an action, then, that action must be sufficiently motivated by desires in this distinguished subset.[33] If it is, then the action is *self-expressive*.

There is, to be sure, much more to any particular deep self theory, but the following general framework is true of them all:

(a) there are some distinguished agential features which constitute a person's deep self

(b) to be morally responsible for an action, that action must bear a distinguished kind of relation to those features

Section 3 – A Supposed Difference

Sripada's observes that it is difficult to adjudicate between deep self and reasons-responsiveness theories because their requirements are so extensionally similar. He says, "Most everywhere we look, control and expression of one's self seem to go hand in hand. A person who has the capacities constitutive of being in control usually exercises these capacities in such a way that her self—or at least some part of it—is expressed in her action. A person with deficits that undermine control usually does something that fails to express her self" Sripada 2017, p. 783). That is, in almost all cases in which a person acts self-expressively, she also acts reasons-responsively, and vice versa, and in almost all cases in which a person acts, but not self-expressively, she also fails to act reasons-responsively, and vice versa.

Sripada's claim here is conditioned on the truth of a (restricted) Humean theory of reasons, according to which a person has a reason to do some thing if and only if it would

[33] Strictly speaking, on the second variant (Frankfurt's view) it is not that these second-order desires motivate the action, but rather that the desires which *do* motivate it are themselves the object of the person's second-order desires.

advance a distinguished subset of her desires—those which constitute her deep self: "*S* has a reason to *A* if *A-ing* advances the desires that constitute *S's* self" (Sripada 2015, p. 251). On this view *there are* only those reasons which correspond in the right way to someone's deep self-constituting desires, and *there are* those reasons only *for* someone with those desires. Assuming that such a theory is correct, it is not difficult to see why the reasons-responsiveness approach is bound to line up with the deep self approach in a very important respect. If the reasons there are for a person in a given situation are exhaustively determined by her deep self, then whether she is reasons-responsive will entirely depend on whether she is disposed to see and act on those deep self-advancing reasons. If she is, then her actions will, by definition, be self-expressive.

Sripada concludes that the difference between deep self and reasons-responsiveness theories is simply that the former require only that a person act self-expressively in the actual scenario under consideration, whereas the latter require that a person would also act self-expressively in a (suitably defined) range of possible worlds. In his own formulation: "Deep self views say moral responsibility requires expressing one's self in action. Reasons-responsiveness views say moral responsibility requires the ability to express one's self across a suitably broad range of alternative scenarios" (Sripada 2015, p. 255). In other words, according to Sripada, deep self theories consider a person to be morally responsible for an action just in case that action actually had the right kind of motivational etiology, whereas reasons-responsiveness theories require that the right kind of motivational etiology would also take place in a number of counterfactual scenarios.

Section 4 – Frankfurt's Willing and Unwilling Addicts

In this section I will present the case which Sripada (2017) believes illustrates this difference and which he uses to argue that we should prefer a deep self account. I will then object that Sripada disagrees with Fischer and Ravizza for the wrong reasons, and that, really, deep self theories like his own are just variants on the general reasons-responsiveness framework.

Harry Frankfurt's (1971) famous cases of the willing and the unwilling addicts center on two individuals who are alike in that they experience literally irresistible desires to use drugs, but who differ in the following respect: whereas the willing addict loves the fact that he experiences and acts on these desires—that is, he wouldn't have it any other way—the unwilling addict

detests this fact about himself and always struggles, if also in vain, to resist his desires to use drugs.[34]

Now, we are supposed to have the intuition that the willing addict is morally responsible for acting on his drug-directed desires, whereas the unwilling addict is not. That intuition is easily explained by a deep self theory, because it is clear that the willing addict acts in a way that expresses his self. We would need to articulate this fact in the terms of one's preferred deep self theory, but the case can be written so that it is true on any one of them that the willing addict's action is self-expressive. Sripada suggests that

> a natural way to fill out the description of the willing addict [is that] he genuinely cares deeply about pleasures of the kind that are received from narcotics. It is of fundamental importance to him that he lives a life imbued throughout with hedonic satisfaction, indeed much more important than the things he surely has to give up in order to maintain his drug-using lifestyle (presumably things such as a fulfilling family life or a flourishing career). (Sripada 2017, p.793)

According to Sripada's care-based view, then, the willing addict acts self-expressively because he is motivated by his fundamental and abiding desire for hedonistic pleasure. The unwilling addict, on the other hand, clearly acts in a way that is not expressive of who he really is. His desire to use drugs is in an important sense obstructing his self-expression. If it were up to him, he would never act on this desire.

Fischer and Ravizza's account, however, is faced with a problem. Recall that a person's mechanism of action is the causal process which led to the action, and that such a mechanism is reasons-responsive just in case certain counterfactuals are true of it. One of those counterfactuals, which I called the *flexibility condition*, states that the mechanism would have produced some alternative action if there had been sufficiently different reasons. If we take a moment to consider just what the willing addict's mechanism of action is, we will see that there is no such possible world, and so the counterfactual expressed in the flexibility condition is not true of it.

The willing addict's mechanism of action clearly includes *both* his normal process of deliberation (about how to advance his deep desires, let us say) *and* his irresistible desire to use drugs. This is because both causally contribute to the action produced, even if that action is

[34] There are other details which Sripada adds to the cases, three of which are worth mentioning in brief: we are told that neither person is morally responsible for acquiring this irresistible desire, that neither person knows that the desire is literally irresistible, and that there is nothing either one of them can ever do to be rid of their desire or of its irresistibility (Sripada 2017, pp. 787-91).

causally overdetermined.[35] And since the desire is irresistible—in the sense that the willing addict's capacity of resistance or self-regulation[36] is too weak to rein it in—it is clear that, holding fixed this mechanism of action, there is no possible world in which he successfully resists the desire.[37] So, if we want to conserve and explain our intuition (should we have it) that persons like Frankfurt's willing addict are morally responsible for acting on literally and permanently irresistible desires, then we should prefer a deep self theory to a reasons-responsiveness theory.

Section 5 – Mechanisms of Action and the Deep Self Alternative

What is it about Fischer's account that gives rise to this problem? Sripada clearly thinks that it is the flexibility condition, which he reads as a kind of leeway requirement. He claims that "[w]hat Fischer has done is only to trade one strong form of access to alternative possibilities […] for a slightly weaker one", whereas "the real point of [Frankfurt-style cases] is that moral responsibility doesn't require access to alternative possibilities at all" (Sripada 2017, p. 811). I believe his diagnosis is mistaken. To see why, consider some things Michael McKenna (2013) has said about reasons-responsiveness theories and, in particular, the role that counterfactuals play in those theories:

> [W]hen an agent acts who is suitably reasons-responsive, the most important factor […] in accounting for her freedom is that the etiology of the act which she actually performed involved springs that were sensitive to reasons. For now, let us think of those springs in terms of the agent herself as the cause of her acts […]. Different reasons, understood as different inputs, would have yielded different outputs, understood as alterations in modes of conduct. And what this shows is that the agent's response to the actual "inputs" played a role that was itself sensitive to, or responsive to, the actual conditions in which the agent acted. To

[35] It is worth noting that Fischer and Ravizza also require the mechanism of action be the "agent's own" (pp. 39, 207-216). Sripada appears to overlook this aspect of the account, but that may be for good reason, since this ownership requirement would run up against the same issue: if the irresistible desire undermines this ownership requirement, then neither the willing nor the unwilling addict will be acting from their own mechanisms of action. As such, the intuitive difference between them will again be lost.

[36] The specific nature of this capacity is not important for present purposes, but see Sripada (2014) and, more recently, Ainslie (2020) for discussions of what it might involve.

[37] The willing addict's mechanism of action is what Fischer calls a *temporally extrinsic* mechanism—a kind of mechanism that is hard-wired to produce a single action. See Fischer & Ravizza (1998), pp. 46-47.

> illustrate, consider a simple example of the sensitivity of a primitive gizmo, a thermostat. Suppose a thermostat is set at 76 degrees (Fahrenheit) and the room the thermostat is in turns out to be 76 degrees. One might wonder if the thermostat's setting accounts for the temperature of the room. After all, it might be disconnected and so merely a fluke that its setting and the room's temperature are 76 degrees. When we learn that the room would come up to 78 were the thermostat set to 78, or would come down to 74 were the thermostat set to 74, and so on for numerous other values high and low of 76, we do not just learn something about the way the thermostat would behave; we also learn about how, in the actual scenario when it is set to 76, it does behave; it plays a certain causal role from reliable and suitably sensitive resources. (McKenna 2013, p. 154)

The truth of the counterfactuals expressed in the flexibility and regularity conditions, then, are meant to determine something about the actual process which led to the action under consideration. Specifically, they are meant to show that the person acted that way *because of* the reasons there were in that situation, just as the room is at its current temperature because of the thermostat's setting.

The flexibility condition, then, is not meant to establish that the person or her mechanism had access to alternative possibilities. Manifestly, the mechanisms in standard Frankfurt-style cases lack such access, due precisely to the counterfactually intervening factors present in those cases. In bracketing those factors, Fischer and Ravizza are not attempting to establish that the mechanism actually could have operated differently:

> On our approach, moral responsibility does not require alternative possibilities. Rather, we have an "actual-sequence" approach to moral responsibility. By this we mean (in part) that one should focus on the properties of the actual sequence in making ascriptions of moral responsibilities. […]
>
> Notice, however, that these "actual-sequence" properties may indeed be *dispositional* or *modal* properties; as such, their proper analysis may involve reference to other possible scenarios or worlds. […] It is important to see that, whereas other possible worlds are relevant to ascertaining whether there is some actually operative dispositional feature (such as weak reasons-responsiveness), such worlds are *not* relevant in virtue of bearing on the question of whether some alternative question is *genuinely accessible* to the agent. (Fischer & Ravizza 1998, p. 53)

Rather, they are attempting to establish that *the way the mechanism actually operated* was sensitive to the reasons at hand. That is an important difference, and one, it seems to me, that Sripada has failed to appreciate.

With this difference underway, we can now see that Frankfurt's willing addict is problematic for Fischer and Ravizza's account *not* because of the flexibility condition, but rather

because of their notion of a mechanism of action. In other words, it is what they choose to *hold fixed* and what they choose to *bracket* in the modal analysis. To see why, consider what would happen if these variables were changed—and changed, moreover, along deep self-theoretic lines. Instead of holding fixed any and all of the person's agential features which were causally relevant to her action, suppose we held fixed

a) the desires constituting her deep self and

b) the action-directed psychological mechanisms motivated by those desires,

and suppose we bracketed

c) any desires which were alien to her deep self.

If these changes were made, how would the modal analysis turn out? It turns out that the flexibility condition will be met for any action that is genuinely self-expressive.

Take, for example, the willing addict: hold fixed his fundamental and abiding desire for hedonistic pleasure, as well as the action-directed psychological mechanisms motivated by that desire. Now, consider what would happen if he were presented with a scenario in which some other, non-drug-involving opportunity promised him more pleasure, and thus a sufficient reason to act otherwise. In such a scenario, surely the willing addict would do otherwise. After all, the only causal contributors to his action will be this desire, the associated psychological mechanisms, and the features of his situation which present an opportunity to advance that desire.

Now compare this with the unwilling addict. The unwilling addict detests drug-involving pursuits, presumably because they are mutually exclusive with the things he deeply cares about. When he uses the drug in the actual scenario, it is because an irresistible, alien desire is undermining his capacity to respond to the reasons there actually are in his situation, given what he cares about. If that desire were bracketed, however, then surely he would not take the drug.

Note that, given this way of conducting the modal analysis, the important difference between the willing and the unwilling addict is not that one satisfies the flexibility condition, whereas the other does not. In fact, both agents, so construed, will exhibit flexibility in action corresponding to changes in their respective situations. If either agent's fundamental desires and associated psychological mechanisms are held fixed, and all alien desires bracketed, there is, as it were, nothing to get in the way of his self-expression. That is, there is nothing which could prevent the agent from responding to each situation in a way that would advance his deep self-constituting desires.

The importance difference between the two agents, then, is that one of the agents is responsive to the reasons there actually are in his situation, whereas the other is not. Conducting the modal analysis thus serves to illustrate what each agent *would do* if he were responsive to reasons, which, when compared to what each agent actually does, reveals that the unwilling addict was not responsive to the reasons in his situation and, therefore, is not morally responsible for taking the drug. In particular, while the unwilling addict correctly recognizes that such reasons exist, due to the irresistible desire he is unable to bring those reasons to bear on his behavior—what Fischer and Ravizza call a "failure of reactivity" (1998, p. 41). In the case of the willing addict, by contrast, there is a clear line running from the reasons there actually are in his situation, through the practical features of his agency, to the action he ultimately performs. He is the one guiding his action in response to the reasons in his situation. He is *in control* of what he is doing in a way that the unwilling addict clearly is not.

Here it is important to remember that the point of disagreement between Fischer and Ravizza's reasons-responsiveness account and Sripada's self-expression account is not that they give different assessments of the unwilling addict's responsibility for what he does. Both conclude that the unwilling addict is not responsible for using the drug. Rather, it is that they give different assessments of the *willing* addict's responsibility for what he does. According to Fischer and Ravizza, both agents are not responsible for the same reason: the actual causal process leading to the agent's action fails the flexibility condition. Their approach is thus unable to explain the intuitive difference between the two agents. However, contrary to initial appearances, this does not support Sripada's conclusion, that *the flexibility condition* is not be a necessary condition of moral responsibility. Rather, it supports the conclusion that the *actual-sequence mechanism* is not the right feature of a person's agency to hold fixed when conducting the modal analysis. After all, the point of the modal analysis is to determine whether, in the actual scenario, the requisite connection exists between the agent's action and the reasons in his situation. Fischer and Ravizza's way of conducting that analysis fails to appreciate the fact that, in the case of the willing addict, such a connection clearly exists, even if his action is simultaneously caused by an irresistible desire.

McKenna actually floats the suggestion that reasons-responsiveness theorists could (and maybe should) avail themselves to the resources of a deep self account in just this way.[38]

> If we hold fixed all of the intrinsic properties constituting Jones, we'd have to include these psychotic-constituting or phobia-constituting properties as well. In such cases, the relevant counterfactuals would come out false as applied to the agents on the proposal currently on offer. […] [One strategy for dealing with such cases] would be to consider whether, when specifying the agent-constituting intrinsic properties that are to be held fixed in counterfactuals […], there is a principled way to rule out those constituting the psychosis or the phobia. The rough idea would be to treat these conditions as in some way alien or distant from those ingredients constituting the agent's identity, or her real self (McKenna 2013, p. 175-176).

Deep self theories are, after all, motivated by the same concern: we cannot consider someone morally responsible for actions caused by *any* of her agential features, since these could include alien and intuitively excusing features like compulsions and hypnotically-induced desires. The notion of a deep self is useful precisely because it gives a principled way of distinguishing between features which are relevant to moral responsibility and those which are not. It is, of course, a separate question—and the one which really matters—whether any of the proposed accounts of deep selves *get those features right*. But I think it is at least clear that Fischer and Ravizza cast too wide a net. Just because an agential feature is causally involved doesn't mean it is relevant.

If that is right, then deep self theories like Sripada's are not really in opposition to what I have identified as the central commitments of the reasons-responsiveness approach. Those commitments are, again, that a person is morally responsible for how she acted in some circumstance only if she was responsive to the reasons in that circumstance, and that, in order to determine whether a person is responsive in this sense, one must conduct a modal analysis comprising three variables:

1. The features which are held fixed across possible worlds
2. The features which are bracketed from our considerations

[38] And he even floats this suggestion as a way of dealing with cases like Frankfurt's willing addict. That is, cases in which it is one of the agent's *intrinsic agential features*—a phobia, a latent psychotic element, a compulsion, or an irresistible desires—as opposed to a feature extrinsic to the agent—another person "waiting in the wings"—that is responsible for ensuring that the person act in a particular way.

3. The counterfactuals (such as the flexibility condition) which must be true if the person is responsive to reasons

Conclusion

The significant upshot of all of this is twofold. First, the problem that Frankfurt's willing addict raises for Fischer and Ravizza's account is not a reason to prefer a deep self theory to a reasons-responsiveness theory. Their account faces this problem because they identify the wrong agential features for their modal analysis, and not because they are committed to the flexibility condition. Relatedly, Sripada seems to think that "control" in this context refers to access to alternative possibilities or a kind of *leeway freedom* at the level of the mechanism of action. As we saw from McKenna's commentary on the role of counterfactuals in the reasons-responsiveness approach, the sort of control that such accounts are after is *source freedom*. They are interested in determining whether the person acted in a way that was actually sensitive to the reasons there were in her situation. Depending on how we understand the nature of reasons, this is just another way of asking whether the agent has acted self-expressively (i.e., because that action will advance her own, self-constituting desires). In defending the broad commitments of the reasons-responsiveness approach against Sripada's charge, then, I have tried to show that these theories are more similar than even Sripada has considered, and that "control" need not be a spooky appeal to alternative possibilities.

The second, related upshot is that source compatibilists—including deep self theorists like Sripada and reasons-responsiveness theorists like Fischer and McKenna—should be more concerned with articulating and defending a principled reason to identify certain agential features as the ones that are relevant for responsibility. Deep self theories are, in my view, more directly attending to this issue in that their accounts seek to make explicit just which agential features constitute the deep self. Fischer's notion of a mechanism of action was, since the beginning, left at an intuitive level—we were supposed to just *know* when two mechanisms of action were the same or not, and also which causal contributors to an action were significant.[39] What reasons-responsiveness and deep self theorists should be engaged in, then, is in direct conversation about

[39] See Fischer & Ravizza (1998), p. 47.

whether the kinds of deep self on offer successfully identify the responsibility-relevant features, or whether something else is more suitable.

In this paper I have argued that deep self accounts are just specified variants of the reasons-responsiveness approach. Chandra Sripada (2015, 2017) has argued that the two approaches are similar, but can be prized apart in one important respect: reasons-responsiveness theories, but not deep self theories, consider *flexible control* to be a necessary condition of moral responsibility. Drawing on Michael McKenna's (2013) discussion of the role that counterfactual conditions play in reasons-responsiveness theories, I have argued that Sripada has misunderstood the flexibility condition in John Fischer and Mark Ravizza's (1998) account, which leads him to disagree with them for the wrong reasons. In my view, Sripada's real point of disagreement is with Fischer and Ravizza's notion of a *mechanism of action*. In fact, I believe that deep self theories like Sripada's are not really in opposition to what I have identified as the central commitments of the reasons-responsiveness approach: namely, that an agent is morally responsible for what she does only if, holding some features of her agency fixed and bracketing the rest, certain counterfactuals are true of the agent, including the flexibility condition.

In my view, the point of substantial disagreement between specific deep self theories, such as Sripada's self-expression view, and specific reasons-responsiveness theories, such as Fischer and Ravizza's, is in *which agential features* they consider relevant for moral responsibility. Deep self theories are specified variants on the reasons-responsiveness approach, in that they more precisely identify which agential features are relevant to our analysis. That is, deep self theories, whether they do this explicitly or implicitly, are in the business of identifying agential features which would be held fixed, and those which would be bracketed, in a modal analysis of responsible agents.

A WILLINGNESS TO BE VULNERABLE: NORM PSYCHOLOGY AND HUMAN-AI INTERACTIONS

Abstract: A distinguishing feature of social robots is that they are designed to be welcomed into roles and relationships which are characterized by a high degree of *vulnerability*: caregiving and assisted living roles, educational roles, and platonic as well as romantic and sexual partnerships. I argue that, as in the human case, our willingness to welcome social robots into these spaces, roles, and relationships should be conditioned on their capacity to understand and live up to *social norms.* Drawing on an interdisciplinary body of research on norm psychology, I explain why this capacity is importantly different from pre-programmed conformity to rules, in that it involves an open-ended, dynamic form of responsiveness to *social corrective feedback,* such as that which humans provide to each other in expressions of praise and blame.

Introduction

Broadly speaking, social robotics refers to artificial intelligence applications that involve some degree of interaction with other agents. Moreover, in many applications social robots are designed to participate in relationships with human users. As discussed in Duffy et al. (1999), Bruce Edmonds (1997) argues that "social intelligence is not merely intelligence plus interaction, but should allow for individual relationships to develop between agents" (Edmonds 1997, p. 1). In this paper I explore an ethically pressing feature of social robots: that they are designed to be welcomed into roles and relationships that are characterized by a high degree of *vulnerability*. Some key examples include the implementation of social robotics in healthcare (van Wynsberghe 2013), especially elderly care (Moro et al. 2019); therapeutic applications (Rabbitt, Kazdin, and Scassellati 2015); domestic roles (Dereshev and Kirk 2017; Young et al. 2009); educational roles (Belpaeme et al. 2018); and relationships of companionship (Wilks 2005) and sexual intimacy (Su et al. 2019).

The vulnerable nature of these roles and relationships is underscored by the selectivity with which we ordinarily welcome someone into them. In ordinary human relationships, when we put ourselves or a loved one under someone's care, or when we open up to a friend or a lover, it is typically because we think we can *trust* that they will be responsive to our or our loved one's expectations about how they should be treated. This connection between vulnerability and trust is,

perhaps, so commonplace that it can be easy to miss, but upon reflection it quickly becomes evident that human sociality is in many ways made possible by our willingness to place ourselves and our dependents at the mercy of others whom we trust. This includes childcare, the sharing and delegation of domestic or professional responsibilities, and participation in interpersonal relationships like friendships and sexual and romantic partnerships, among many others.[40] Although some social robotics applications may be more innocuous, such as a robot designed to keep the home or workplace neat and organized, clearly many of the purposes for which social robots are being designed involve forms of interpersonal interactions where the risk of bodily, psychological, and emotional harm is a distinct possibility.

The guiding question of the paper is this: Should we welcome social robots into these roles and relationships of vulnerability, and if so, on what conditions? By posing the question in this way, I follow what Dorna Behdadi and Christian Munthe (2020) call a "normative approach" to artificial agency. They argue that the debate surrounding artificial agents, and artificial moral agents in particular, "should be redirected to address outright normative ethical questions. Specifically, the questions of how and to what extent artificial entities *should* be involved in human practices where we normally assume moral agency and responsibility" (Behdadi and Munthe 2020, p. 214). Instead of debating about the necessary and sufficient conditions for moral, social, or responsible agency, or for "genuine" interpersonal relationships, I will attend more directly to the practical question of whether we *should* welcome social robots into roles and relationships which characteristically involve a high degree of bodily, psychological, and emotional vulnerability on the part of human users. As such, my approach also shares affinities with Vincent et al. (2015), who explore human perspectives on social robots. These are perspectives, namely, that focus on the social practices of humans and "how these practices interact with […] social robots", such that "[h]uman needs, senses, emotions, desires and attitudes all foreground the discourse" (Vincent et al. 2015, p. 2).

Whereas a complete answer to this paper's guiding question would turn on an incredibly expansive range of factors, my more modest aim is to show that an adequate answer to that question will have to countenance the ways in which humans *already do* accept and navigate the vulnerability that characterizes their relationships with one another. In particular, I argue that we

[40] For a related discussion about the importance of safety and security for interactions between social robots and the elderly, see Akalin, Kristoffersson, and Loutfi (2019).

should welcome social robots into these relationships only if they are responsive to specifically *normative* forms of social interaction, exemplified by praise and blame, which humans use to cue each other in to, and hold each other accountable to, their interpersonal expectations.

I begin in Section One by considering the significance that human beings attach to certain social expectations in their ordinary interpersonal relationships, and why we should look to these relationships as models for the expectations we would (and should) have of social robots. In Section Two I give a broad overview of the psychological capacity which enables humans to detect, internalize, and be motivated by these expectations—or, *social norms*—and I emphasize the role that a variety of *normative responses* play in communicating and enforcing these expectations. Section Three argues that, just as in the human case, we should expect more from social robots than mere rule following. What is required is a higher-order capacity to *update* behavior in response to the corrective feedback which users provide through normative responses like praise and blame. On the picture I present, our willingness to be vulnerable with social robots should be conditioned on a specific type of social-normative learning.

Section 1 – Human Relationships as Model

In P. F. Strawson's seminal essay, "Freedom and Resentment" (1962), he urged philosophers not to forget what ordinary human relationships are actually like, particularly when theorizing about things like moral responsibility. While my concern in this paper is not with the nature of moral responsibility and does not try to answer anything as metaphysically laden as the question of whether social robots could be genuinely responsible for what they do, I do think that Strawson's emphasis on the ordinary nature of human relationships is well placed in a discussion about our willingness to welcome social robots into the roles and relationships for which they've been designed. As the verisimilitude between human-human and human-AI interactions increases[41], so does the likelihood that humans will be disposed to interpret and interact with

[41] This is not to suggest, however, that verisimilitude always is or should be the goal for social robots, although some degree of at least *functional* similarity (as opposed to bodily similarity, for instance) does seem necessary for relationships of companionship and caretaking. For a broader discussion of non-humanoid approaches to social robots, see Mokhtar (2009).

social robots in ways that closely resemble their interpretations of and interactions with human beings.[42] As Mark Coeckelbergh (2010) writes,

> if robots were sufficiently advanced—that is, if they managed to imitate subjectivity and consciousness in a sufficiently convincing way—they too could become the quasi-others that matter to us in virtue of their appearances. As emotional and social beings, we would come to care about how we would appear to robots—about what robots would 'feel' and 'think' about us. (Coeckelbergh 2010, pp. 238-239)

Similarly, Höflich and El Bayed (2015) observe that, "[w]hen people engage in a relationship with a robot, they engage in a dyadic relationship on an 'as-if' basis of imagination; as if the robot was a social being (Höflich & El Bayed 2015, p. 47). This is why it is less pivotal than traditionally assumed whether social robots can "genuinely" relate to human beings in the ways humans relate to one another, particularly insofar as these relationships might be thought to require emotional states and conscious experiences. What is at stake is a much more pragmatic concern: to the extent that the *humans* in human-AI interactions will have experiences similar to those they already have with other humans, to include psychological and emotional harms, what should we do? Should we, collectively and individually, welcome social robots into roles and relationships which involve this sort of vulnerability, and if so, on what conditions?

I believe that an adequate answer to this question must countenance the same ordinary social fact which Strawson then emphasized: that human beings exhibit a great deal of concern for other people's attitudes and intentions, as evidenced by the *reactive* attitudes they are disposed to feel when other people treat them with good or ill will.

> We should think of the many different kinds of relationships which we can have with other people—as sharers of common interests; as members of the same family; as colleagues; as friends; as lovers; as chance parties to an enormous range of transactions and encounters. Then we should think, in each of these connections in turn, and in others, of the kind of importance we attach to the attitudes and intentions toward us of those who stand in these relationships to us, and of the kinds of *reactive* attitudes and feelings to which we ourselves are prone. In general, we demand some degree of goodwill or regard on the part of those

[42] Studies on human-robot interactions are not only suggestive of this effect, but tend to be premised, at least implicitly, on the goal of increasing human acceptance of artificial agents. See for example Maartje, de Graaf, Allouch, and van Dijk (2019) and de Graaf, Allouch, and Klamer (2015). For a review, see Katz, Halpern, and Crocker (2015).

> who stand in these relationships to us, though the forms we require it to take vary widely in different connections. (Strawson 1962, pp. 6-7)

This concern for others' attitudes toward us, and the reactive attitudes which we are prone to experience *about* and *in response to* those attitudes—such as resentment, gratitude, contempt, shame, and forgiveness, among many others—Strawson took these to reflect a basic demand to be treated with goodwill. That basic demand can take on a variety of different forms depending on the specific relationship in question, but most ordinary relationships appear to involve some form of the general expectation that others will treat us as we think we should be treated.

While some of the forms which this basic demand for good will can take may be more universal in scope than others, such as a demand not to be physically harmed, I want to reflect for a moment on the more contingent and contextual forms it can adopt, as when we choose to enter into relationships of a more selective nature. Consider, for example, the sort of expectations we have about ordinary friendships. A friend is, among other things, a person with whom we are willing to share certain private details of our lives. We might share, for example, something embarrassing we did when we were drunk, or an argument we had with a coworker, or our crippling fear that we will never amount to anything. There are many reasons we choose to open ourselves up to a friend in these ways, from reassurance to commiseration to advice, all of which contribute to the significance and the value that such relationships have for us. But an implicit condition on our doing so is that the friend will not turn around and disclose our secrets to the world.

Not only is trustworthiness a condition of many actual friendships, it is also one which most people have at some time had the unfortunate occasion to see violated. When such violations occur, they tend to inspire feelings of sadness, anger, betrayal, resentment, and blame. This is because many such ordinary interpersonal relationships involve—by design, so to speak—some measure of risk and vulnerability, and so our willingness to welcome someone into them rightly tends to be premised on a wide range of overlapping and crosscutting expectations, which allow us to be vulnerable with each other in ways that we all sometimes need to be, but cannot be with just anyone. When these expectations are violated, our relationship with that person is impaired, and in response it is not only likely but apt that we will experience feelings and attitudes which reflect that impairment (Scanlon 1998; Smith 2013). We expect our partners in interpersonal relationships, from friends to lovers, to treat our concerns with the seriousness

they deserve, and, more generally, to treat *us* as persons whose concerns deserve to be treated seriously. When we believe that someone with whom we've taken the risk of friendship or of intimacy has failed to treat us with this respect and this good will, the significance it can have for us is a normative consideration par excellence. It calls into question the very relationship in question, by revealing that that person "does not have the attitudes, dispositions, and intentions that are (ideally) constitutive of a relationship of friendship" (Smith 2013, p. 37).

When I say that a distinguishing feature of social robots is that they are designed to be welcomed into spaces, roles, and relationships which are characterized by a high degree of vulnerability, this is the sort of vulnerability I primarily have in mind. Once we appreciate the significance that human-AI relationships are likely to have *for humans,* it becomes surprisingly unimportant whether social robots do or ever could meet the kind of metaphysical conditions which putatively suffice for, say, responsible agency or friendship or romantic love. What matters are the potential risks which come with welcoming social robots into roles and relationships whose significance for humans centers on a willingness to be vulnerable, and which are therefore ordinarily premised on expectations like those I've discussed in the particular case of friendship. It is not incidental to the ethical design space of a social robot, especially one which is intended to approximate the psychological profile of a companion, caregiver, or lover, that human users will attach significance to the feelings and attitudes which they believe the social robot has toward them. This is why human-human interactions can serve as a model for what we would and should expect of social robots that are designed to occupy similar roles and to participate in similar relationships: What humans *would* expect of social robots—where this is a purely descriptive claim about a psychological tendency of humans—tells us something important about the normative consequences which human-AI relationships would have *for humans*. This is what allows us to draw a prescriptive conclusion about what we *should* expect of social robots: namely, that they be able to understand and live up to the expectations of the individuals with whom they interact.

Section 2 – Human Norm Psychology

In this section I discuss how, i.e., psychologically, human beings are able to understand, live up to, and hold each other to the sort of expectations glossed in the previous section. The Strawsonian idea, that ordinary human relationships are structured by expectations about how it

is appropriate to treat one another, has in recent years found a more precise formulation in the rapidly expanding and interdisciplinary body of research on *norms* and *normative cognition*. That research suggests that humans possess a suite of psychological mechanisms, referred to collectively as a *norm system,* which enable them to detect, understand, internalize, and be motivated by the norms in their local communities (Kelly 2020, forthcoming; Sripada & Stich 2007). Considering the basic contours of the norm system will take us another step toward answering this paper's guiding question. In particular, it reveals the highly dynamic form of social responsiveness on which those norms depend.

Norms are informal rules or standards which govern what is allowed, disallowed, expected, and required for different roles, relationships, and activities. To say they are informal is just to mark the fact that, although norms are sometimes codified in formal laws, they are often simply implicit in people's attitudes and behavior. So understood, the range of things which could, in principle, be governed by norms is quite expansive. Some of the more obvious examples are norms of a putatively *moral* nature, such as a norm which forbids intentional killing except in self-defense, but there are also norms which govern the comparatively mundane, such as a norm which stipulates how close it is appropriate to stand to someone during a conversation.[43] Norms also exhibit a high degree of cultural variability. As is well known, what is appropriate or even expected in one community can be distasteful or condemnable in another.[44] But the presence of some system of norms appears to be culturally universal. Moreover, part of what it means to say that there is a norm in a given community is that members of that community not only tend to conform to the prescription of the norm if it applies to themselves, but also tend to *enforce* conformity on others, particularly in response to norm violations.

We now know that as early as three to five years of age humans begin to exhibit an understanding of different norms (Turiel 1983; Smetana 1993; Nucci 2001) and a facility with rule-based reasoning tasks (Cummins 1996; Beller 2010). They also exhibit a heightened sensitivity to features of their local environments, particularly the behavior of other humans,

[43] For a comprehensive review of the different taxonomies which have been proposed for norms, and for an argument in favor of a pluralistic approach, see O'Neill (2017).

[44] However, it should be said that the mere presence of a norm does not speak to its legitimacy. Norms can be unjust and oppressive, such as, to name just one example, norms which forbid participation in certain social activities on the basis of someone's gender.

which indicate the presence of norms. Once the presence of a norm is detected, preschoolers will quickly infer the rule which that norm corresponds to (Rakoczy, Warneken, & Tomasello 2008), even without formal instruction (Schmidt, Rakoczy, & Tomasello 2011), and go on to enforce the norms they learn on others (Josephs et al. 2016).

The ability of humans to reliably track norm-relevant features of their environments and translate this sort of input into norm-guided behaviors suggests that they possess a distinctive capacity for norm-guided cognition and behavior. According to a standard approach in psychology and the cognitive sciences known as *homuncular functionalism* (Lycan 1990), complex capacities such as the human capacity for norms can be explained by decomposing them into the simpler mechanisms responsible for carrying out the various tasks associated with that capacity.[45] Chandra Sripada and Stephen Stich (2007) give one such model of the human norm system, which comprises two psychological mechanisms:

1. A norm acquisition mechanism, which is responsible for
    a. Detecting behavioral cues which are indicative of the presence of a norm
    b. Inferring the prescriptive or proscriptive content of that norm, and
    c. Communicating this information to a norm execution mechanism
2. A norm execution mechanism, which is responsible for
    a. Encoding the normative information communicated to it by the norm acquisition mechanism in a norm database
    b. Detecting environmental cues which indicate that a norm applies in the present situation, and to whom it applies
    c. Motivating conformity to norms which apply to oneself, and
    d. Motivating punishment of norm violators

Notice that, in addition to detecting and inferring the content of norms, the norm acquisition mechanism passes this information along to a norm execution mechanism which encodes the information it receives into a norm database. Another way in which this process is sometimes described is as a process of *norm internalization* (Bicchieri, Muldoon, & Sontuoso 2018; Kelly

[45] Eventually this process of decomposition would reach levels of description which are purely mechanistic and which are suitable to be realized in the neurological and physiological structures of the individual's brain and body (Marr 1982; Dennett 1981, p. 80)—or the hardware of an AI—but I will be concerned only with the "computational" or task level of analysis.

forthcoming, 2020; Kelly & Davis 2018; Fehr & Falk 2002). While the exact nature of norm internalization is the subject of some debate, it is generally understood in terms of its motivational effects.[46] Once a norm is internalized, individuals become highly motivated to conform to the norm, to evaluate their own and others' behavior against it, and to enforce conformity on those to whom the norm applies. That is, they become highly motivated to engage in a variety of enforcing behaviors, such as reward, punishment, praise, and blame.

There a number of proposals for how humans would have come into possession of such a norm system, some which attribute greater significance to evolutionary selection pressures, and others which play up the role of enculturation. According to the former, the norm system involved in response to selection pressures present in the ancestral environments of humans which favored individuals who were better able learn from and conform to the behavior of conspecifics, especially in connection with in-group cooperation and the inheritance of accumulated cultural knowledge and technology (Gintis, Bowles, Boyd, and Fehr 2005; Boyd & Richerson 2005a; N. Henrich & J. Henrich 2007; Tomasello 2009). On this sort of view, the norm system came to be part of the innate genetic endowment of all human beings due to their shared evolutionary history. Other proposals lay less emphasis on evolutionary selection pressures, and more emphasis on the enculturation process. On this alternative sort of view, humans develop a capacity to recognize and be motivated by norms in something like the way they develop a capacity to read and write—by using domain general learning capacities to acquire skills which are practiced and taught to them by other members of their community (Heyes 2018). What explains the cultural universality of norms on this sort of account, then, would be the universal presence of that social technology in all human cultures.

Importantly, even if the norm system is innate in the way described by evolutionary accounts, the *specific norms* which an individual internalizes will entirely depend on the presence of norms in the community in which she is enculturated (Boyd & Richerson 2005b; Sripada & Stich 2007; Chudek & Henrich 2011; Kelly & Davis 2018). That is to say, while all

[46] Perhaps the nature of this motivation is purely instrumental, in that it descends from the individual's desire not to be punished (Bicchieri, Muldoon, and Sontuoso 2018). Alternatively, some theorists believe that internalized norms motivate individuals intrinsically (Kelly forthcoming, 2020; Kelly and Davis 2018). That is to say, they motivate compliance and enforcement not in order to avoid averse consequences, but rather for the simple reason that they interpret violations of the norm as *wrong,* full stop, and so they would conform to its prescriptions even when the threat of punishment is absent.

developmentally normal humans will come to have this general capacity to detect, understand, internalize, and be motivated by norms, the norms they are actually motivated to conform to and enforce will be those which they learn through social interactions and cultural inheritance (cf. Tooby & Cosmides 1992). This underscores the cultural variability of norms, as well as their developmental nature: humans depend for their facility with specific norms on the opportunities they've had to internalize those norms through her interactions with others.

In this section I detailed the human capacity to detect, understand, internalize, and be motivated by norms, including the kind of social norms I discussed in Section One in connected with friendship. This *norm system* is what explains, at a proximate psychological level, the significance humans attribute to certain interpersonal expectations. In the next section I connect this account back to the guiding question about human-AI relationships, and I argue that our willingness to be vulnerable with social robots should be conditioned on their possession of a similar capacity to recognize and respond to normative feedback.

Section 3 – Designing a Normatively Capable Robot

In Section One I argued that our willingness to be vulnerable in ordinary interpersonal relationships is premised on certain expectations, and in Section Two I reviewed a body of research which suggests that, in human beings, the capacity to identify, navigate, and enforce these expectations is due to an evolved or culturally acquired capacity for normative cognition. This corresponds, roughly, with Edmonds' (1997) claim that an important aspect of social intelligence is "the development of rules to structure social interaction—either formally or informally (e.g. emergent social norms, or formal procedure)" (Edmonds 1997, p. 4).

A preliminary answer to the guiding question of this paper—again, whether we should welcome social robots into roles and relationships that are characterized by vulnerability—would thus appear to be: *yes,* so long as social robots are designed to follow the rules.

Notice that, so articulated, the solution is to make social robots *predictable* vis-à-vis our normative expectations—i.e., to program them in such a way that they will consistently follow a catalogue of social norms and will not suddenly "go off the rails." This putative solution therefore immediately raises the further questions of *what the norms should be* for a given social robot, considering the role or relationship it is designed for, and *who should be responsible for determining the contents of those norms* (i.e., should a software engineer be the one to decide

how my robot friends will treat me?).[47] Not only does this putative solution appear to raise more questions than it answers, it also, I believe, fails to appreciate the dynamic nature of the phenomenon in question. In this section I describe the importance that *reactive exchanges* (McGeer 2015) have for ordinary interpersonal relationships, and I argue that, if social robots are to be welcomed into roles and relationships of vulnerability, then they will need to be designed so as to have a capacity, like the human capacity for normative cognition, that is dynamically responsive to specific social cues.

As we saw in Section Two, once someone has internalized a norm, she is thereby highly motivated to enforce conformity on others. That is, she becomes highly motivated to engage in a variety of what I will call *normative responses*. Among these normative responses are precisely the sort of reactive attitudes which Strawson emphasized and which have since taken center stage in the responsibility literature under the umbrella of praise and blame.[48] Normative responses can range from the relatively subtle, as when someone shakes her head in dismay or gives someone the "stink eye," to overt and emotionally charged confrontation. The regulative effects that these normative responses have on humans is key to understanding the way in which norms become internalized by individuals and eventually stabilize in the wider population (Boyd & Richerson 1992). Normative responses like praise and blame thus serve an indispensable social function: they communicate to offenders that their behavior is unacceptable, thereby reinforcing the norm's significance for their future behavior, and they simultaneously signal the presence of the norm to third parties.

While there is disagreement among responsibility theorists about the significance we should attribute to the communicative and regulative effects of normative responses like praise and blame, particularly with respect to questions of justification (Scanlon 1998, pp. 275-277; cf. Vargas 2013), most theorists acknowledge that these responses do characteristically have these effects. Moreover, when we respond to others in one of these ways, we generally expect that our response *will be responded to, in turn*. That is, when an individual violates one of the norms

[47] Note that many of these ethical difficulties mirror those raised by self-driving cars. For a review, see Nyholm (2018).

[48] Responsibility theorists are not all in agreement about the full range of normative responses which are relevant for moral responsibility, and an increasingly popular position is that there are distinct types of responses, each which its own agency conditions (Watson 1996; Shoemaker 2015, 2011; cf. Smith 2015).

which govern our relationship with her, we tend to engage her in what Victoria McGeer (2015) has called a *reactive exchange.*

> We are disposed by nature and nurture to experience what Strawson called "reactive attitudes" when we take others to fail or to succeed in living up to what moral norms require of them – attitudes like resentment and indignation, gratitude and approval. And we are equally disposed to be responsive to those reactive attitudes when others display them to us – not simply by experiencing self-reactive attitudes in turn (e.g. guilt, shame, remorse, pride, self-satisfaction, etc.), but also *by acting in ways those attitudes make appropriate*.. So, for instance, in the face of others' resentment or indignation, we are not only inclined to feel badly, we spontaneously offer excuses, justifications, explanations, and sometimes even apologies when we take our action to be unjustified […]. [S]incere remorse and genuine apology will often pave the way to mollification and "forgiveness", where forgiveness should be understood (at least in part) as a reactive attitude *acknowledging the offender's understanding of what our shared norms require* and (re)commitment to respecting them. (McGeer 2015, pp. 272-273; my emphasis)

Reactive exchanges are, in effect, attempts to *repair* the damages which our relationships incur when others violate certain normative expectations. In Section One I stressed that, in order to answer the question of whether social robots should be welcomed into roles and relationships which are characterized by vulnerability, we need to consider what humans will likely experience in these relationships. The fact that we are willing to, and even insist on, engaging in these exchanges with one another reveals that our expectations of one another go beyond the first-order prescriptions of various social norms. They also include expectations about how it is appropriate to respond to one another when normative expectations fail to be met. When we blame someone, for example, the significance which that interaction has for us is not just that she has violated one of the norms of our relationship; it is also lies in our expectation that she will respond to our blame in ways that acknowledge the legitimacy of our redress—that she will act "in ways those attitudes make appropriate" (McGeer 2015, p. 272).

It is tempting to insist that reactive exchanges such as these would be unnecessary if social robots could be designed so as to never deviate from our normative expectations of them. But besides being rather unlikely, given the high degree of cultural and individual variability in our expectations, such a solution would fail to appreciate the significance humans attribute to these reactive exchanges themselves.

The kind of trust which I have argued is a condition of our willingness to be vulnerable with someone is, in human-human relationships, the product of a history of disagreements and

resolutions. It is, moreover, premised on the background condition that raising such disputes and seeking subsequent resolution is an ongoing possibility. A good friend is not merely a rule-follower; she is also someone who will navigate and negotiate the terms of her relationship with us as that relationship, and as the individuals in it, evolve over time. And when we decide to enter into such a relationship with someone, we do not simply rely on our ability to *predict* the other person's behavior. That is to say, we do not simply attempt to foretell whether a given individual is disposed to follow the norms of the relationship or role in question. We also, and perhaps primarily, attempt to gauge whether this individual will be *responsive to the feedback we provide* as our own understanding of those norms becomes more refined. Because *we* are continually learning at a normative level, so much the machines. There is, in other words, no static list of first-order normative expectations which could be used to program a social robot, even one that is customized for an individual user. If there is any constant to our normative expectations of others, it is that they will remain open to the possibility of normative revision through reactive exchange.[49]

What we should expect from social robots, then, is already well modeled by the psychological capacity described in Section Two. The principal way in which human agents actually learn about, internalize, and thereby become motivated to follow norms is precisely through being held accountable by others. That is to say, their capacity to understand and live up to social norms is mediated by a wider system of social practices—the ways others engage them and hold them accountable—which functions to cue them into the presence of those norms and the roles, relationships and contexts to which they apply. The human norm system *works* only because it is situated in and sensitive to the social environment, and especially to the attitudes of praise and blame which other people exhibit in response to norm violations. An artificial norm system could operate by largely similar principles. As in the human case, this capacity would not be a simple, first-order disposition to conform to pre-programmed rules; it would involve an open-ended, dynamic form of responsiveness to social corrective feedback, such as that which

[49] This roughly corresponds to the distinction identified in Edmonds (1998) between agents that merely "apply their intelligence" in their interactions with others, such as intelligence about how to conduct trades, and agents that apply their intelligence "to the *process* of relating to the others" (Edmonds 1998, pp. 684-85)—i.e., agents with the capacity to recognize other agents as individuals with desiderative and epistemic states and update those representations in response to those agents' behaviors.

humans provide to each other through reactive attitudes and expressions of praise and blame. Not only does this discharge the difficult task of determining which rules social robots should follow, in that their facility with specific norms will be very closely connected to their history of interactions with their social partner or partners; but it also enables social robots to engage in the kind of dialogical process which, in human relationships, makes vulnerability a risk worth taking.

Conclusion

In this paper I have argued that social robots should be welcomed into roles and relationships of vulnerability only if they are responsive to social corrective feedback, exemplified by praise and blame, which serves to cue them into the evolving social expectations of human users. I started from the observation that many social robotics applications are designed for roles and relationships that are characterized by bodily, psychological, and emotional vulnerability. Drawing on resources from Strawson (1962), I argued that the *humans* in human-AI interactions will be psychologically disposed to attribute special significance to the attitudes they believe their robot counterparts have of them, especially as those interactions come to more closely resemble ordinary human relationships. Then I refined this Strawsonian idea by reviewing an emerging literature on the psychology of normative cognition, and I noted that the human capacity for norms works through its reponsiveness to social feedback that cues individuals into the presence of norms, which they then become highly motivated to follow and enforce. Finally, I suggested, in concert with McGeer (2015), that the significance which humans attribute to social norms extends beyond mere rule following. That is, humans regularly engage one another in *reactive exchanges* that are oriented toward repairing damages to their relationships with others. It is, I argued, an individual's openness to engaging in this sort of exchange that serves as the basic background condition of trusting relationships. As such, social robots should be equipped with a similar capacity to recognize and respond appropriately to social feedback such as praise and blame.

ADDICTION IS NOT DIMINISHED PERSONHOOD: THE ROLE OF DEPENDENCIES IN SELF-GOVERNANCE

Abstract: Individuals with addictions continue to drink, smoke, use drugs, or gamble even against their own better judgment. If that were the sole explanatory target of a theory of addiction, then it would only be natural for the theory to place emphasis on a process, neurobiological or otherwise, whereby the individual's capacity for self-control or self-governance is overpowered, hijacked, bypassed, or diminished. While a diminished capacity to self-govern is no doubt central to many cases of addiction, and is probably a factor in most cases of addiction at some point, in this paper I argue that addiction can also be an expression of self-governance. That is to say, sometimes the reason an individual with addiction engages in addictive behavior is not that she has failed to govern herself according to her better judgment, but is rather that she has successfully governed herself according to a particular vision of who she is, what she values, and what she is capable of. Specifically, I argue that certain features of an individual's personal narrative, such as a belief about oneself as dependent upon something, lead the individual to govern herself according to personal rules that prescribe addictive behavior, rather than proscribe it.

Introduction

Individuals with addictions[50] continue to drink, smoke, use drugs, or gamble even against their own better judgment. If that were the sole explanatory target of a theory of addiction, then it would only be natural for the theory to place emphasis on a process, neurobiological or otherwise, whereby the individual's capacity for self-control or self-governance is overpowered, hijacked, bypassed, or diminished. Notably, this explanatory framework is common to both disease and choice models of addiction. Where theories of addiction as a neurobiological disease emphasize persistent neuroadaptations that entrain automatic and even compulsive motor responses to associated cues (Hyman 2005; Robinson & Berridge 2000; Leshner 1997), choice-based theories point to general features of voluntary choice that incline humans toward impulsivity, which are exacerbated by addictive drugs, especially in individuals with comorbid mental disorders (Heyman 2009; Ainslie 2001). The common assumption is that what needs explaining in

[50] I have chosen "individual with addiction" as a replacement for "addict". "Person with a substance use disorder" is simply too long and acronyms (e.g., "PWSUDs") too depersonalizing. I'll also be making frequent references to personhood, personal narratives, and personal rules, so "individual" added a needed bit of variety. It may not be ideal, but I think it is an acceptable trade-off. See Pickard (2020), pp. 3, 16-17, for a discussion of this issue.

addiction is the fact that individuals with addiction fail to govern themselves in ways that reflect their own better judgment, or in ways that express who they really are and what they really value. The tendency may even be taken to reflect a bias in the literature toward *recovering* addicts—individuals who, having rejected their addiction, struggle to change their behavior accordingly.

While a diminished capacity to self-govern is no doubt central to many cases of addiction, and is probably a factor in most cases of addiction at some point, we should be careful not to mistake even a big piece of addiction for the whole puzzle. In this paper I resist this common narrative by arguing that addiction can also be an expression of self-governance. By this I mean to say that, sometimes, the reason an individual with addiction continues engaging in addictive behavior is not that she has failed to govern herself according to her better judgment, but is rather that she has successfully governed herself according to a particular vision of who she is, what she values, and what she is capable of. Specifically, I argue that certain features of an individual's personal narrative, such as (but not limited to) a belief about oneself as dependent upon something, may lead the individual to govern herself according to personal rules that prescribe addictive behavior, rather than proscribe it. My thesis is thus part of an alternative approach to addiction, which maintains that it is simply not possible to "explain many cases of addiction without recognizing the *value* of drugs" and without recognizing the "sense of self and social identity" that often accompanies addiction (Pickard 2020, p. 2).

I present my argument as a critical response to Don Ross's *Diminished Personhood Account of Addiction* (DPA)[51], who gives one of the most compelling cases for thinking that addiction is a form of diminished capacity for self-governance. Part of what makes the DPA so compelling is the facility with which it embraces both the neurobiological and the choice-theoretic aspects of addiction. In short, Ross argues that individuals with addiction come to see themselves as incapable of following their own personal rules, due to neuroadaptations that result from addictive behavior and undermine their capacity to enforce those rules. After presenting the DPA in Section One, in Section Two I show that individuals with addiction may interpret their

[51] I chose this name for Ross's view based on an unpublished (but publically accessible) manuscript in which he argues that addicts and elephants are two varieties of diminished persons. The core elements of the 2017 view are reproduced in Ross (2020), in which Ross uses the DPA to frame his argument that addiction is "not a disorder 'of the brain', in the sense that no part of the addict's brain is directly damaged *by addictive neuroadaptation* […] and no part of the brain is malfunctioning with respect to its evolutionary proper function" (p. 6).

relationship to their addiction rather differently than Ross envisions. Namely, they may come to represent themselves as dependent upon the substance or activity in question in order to achieve the things they value. In this way, addictive behavior may come to have instrumental value that leads individuals to adopt personal rules *for*, rather than *against*, that behavior. In Section Three I conclude by addressing a number of objections.

Section 1 – The Diminished Personhood Account of Addiction

The *Diminished Personhood Account of Addiction* (DPA) argues that addiction is a variety of diminished personhood, where personhood is defined in terms of a capacity for self-governance. In order to fully appreciate this claim, it will be necessary to present the Self-Governance Theory of Personhood (SGP), on which it is based. But first, I want to get the basic shape of the DPA on the table, so that the connection to addiction will be clearer throughout. The DPA analyzes addiction as a cycle with three, mutually reinforcing stages, depicted in Figure 3 below.

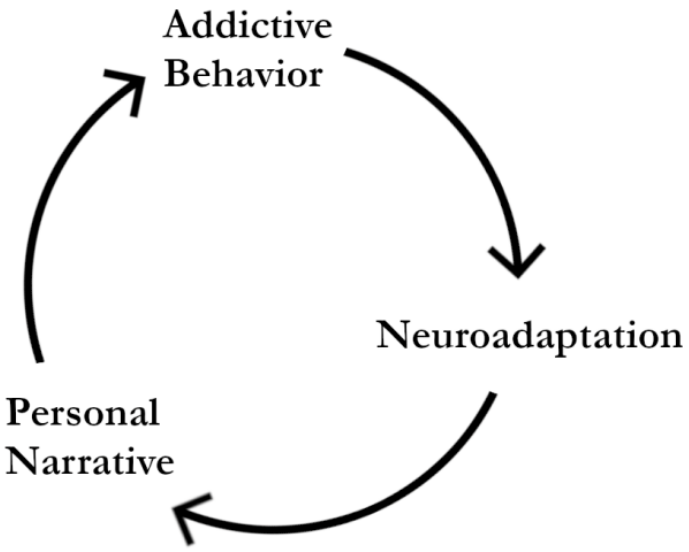

Figure 3. The figure above depicts the cycle of addiction according to the Diminished Personhood Account of Addiction (DPA).

First is the addictive behavior itself, such as gambling or the consumption of certain drugs. Individuals who do not yet have an addiction may engage in these behaviors for a number of reasons, such as anticipation of positive changes to mental state (e.g., euphoria, relief of pain).

Second are the neuroadaptations that result from addictive behavior. Whether these changes constitute a neurobiological disease is a highly contested issue, and one that I won't attend to in this paper. But it appears safe to say that such changes do occur as a result of addictive behavior, and that they occur, moreover, in the reward learning system of the brain. Beyond the transient consummatory or hedonic pleasure that results from taking drugs, drinking alcohol, or getting a big payout at the slots, there are also longer lasting changes to areas of the brain responsible for the prediction and valuation of rewards, motivation to seek out rewards, and attention to reward-associated cues. Most theorists of addiction agree that these changes make it much more difficult to resist addiction-related desires when they arise, and some argue that addictive behavior becomes compulsive once the connection between that behavior and certain cues becomes highly sensitized (Robinson & Berridge 2000). For the DPA, the significance of these neuroadaptations is the effect they have on an individual's capacity to *self-govern.* Specifically, they make it less likely (though not necessarily impossible) that the individual will effectively enforce *personal rules* on her behavior, such as a rule of abstinence, or a rule which allows the behavior under restrictive conditions (e.g., only while socializing). Finally, in the third stage of the cycle, the individual interprets her failure to enforce personal rules as evidence that she is incapable of self-regulation, and she updates her *personal narrative* to reflect this. Ross illustrates the process as follows:

> The gambler, for example, promises herself as she sits down at the slot machine that she will stop playing when the cash stake she is carrying in her purse is exhausted. But then, if she is an addict or on the path to addiction, when that point comes she is likely to find herself reaching for her credit card. Her failed prediction on this occasion undermines her confidence in the value of setting personal rules in the first place. This dimunition [sic] of confidence will tend to exert causal influence on her behavior, which in turn affects her brain, as her repeated play strengthens phasic dopamine response, attenuates GABAnergic interference with signals to motor neurons, and ultimately generates cravings when she isn't gambling. *As the dominant narrative about herself that she will tell*, her status as a person who cannot sustain self-regulation will influence her level of investment in alternative possible life courses, and infect the confidence of those who might try to help her. Addicts are thus pulled into self-reinforcing loops, as diminishing personhood leads to more gambling, which promotes addictive neuroadaptation, which in turn makes full personhood harder to credibly maintain. Most addicts go through phases, which can last for months or years, *in which they give up formulating personal rules altogether*. (Ross 2017, p. 9, emphasis mine)

According to the DPA, when such an individual fails to enforce personal rules, the consequence is not just the immediate one that she will proceed to engage in addictive behavior, resulting in further neuroadaptation. For as she repeatedly fails to impose these rules, she will also begin to see herself in a new light. Namely, she will begin to represent herself, in the narrative she tells herself and others[52], as the kind of person who is incapable of imposing personal rules in general. As such, she will no longer see value in attempting to adopt and enforce more effective personal rules, leading her to acquiesce in the face of addiction-related desires going forward. What is especially interesting about the DPA is the connection it thus draws between the neurobiological aspects of addiction and a narrative-driven capacity to make decisions according to personal rules. As I'll now turn to discussing, this latter capacity is central to the DPA's claim that addiction is a form of diminished *personhood.*

To be sure, Ross is not the first to speak of addiction in connection with personhood. In the highly influential essay, "Freedom of the Will and the Concept of a Person" (1971), Harry Frankfurt contrasts three individuals who all experience irresistible desires to use drugs, but who differ with respect to how they view these desires. First is the "unwilling addict". The unwilling addict detests the fact that he experiences these desires, and he tries mightily but to no avail to resist them. In Frankfurt's terminology, the unwilling addict has a *second-order* desire to avoid acting on his *first-order* desire to use drugs. Next is the "wanton". The wanton has no opinion whatever about his desires. He simply never considers whether his desire to take the drug, or any other desire, is itself desirable. He thus has no second-order desires at all. Last is the "willing addict". The willing addict is "altogether delighted" by the fact that he experiences and acts on his desires to use drugs and "wouldn't have it any other way" (Frankfurt 1971, p. 19). He is like the unwilling addict, in that he has a second-order desire that is about his first-order desire to use the drug. But unlike the unwilling addict, the desire he wants to ultimately act on is the desire to use.

Frankfurt's essay serves well as a launching off point for a discussion of personhood, because it quickly and suggestively illustrates the difference between two forms of agency, one more "personal" than the other. In one form, exemplified by the willing and the unwilling addicts, the agent is "at a remove" from her own first-order desires. This is an agent who does not merely

[52] For a related discussion of the significance of narrative identity in human life, see MacAdams (2019).

act from her desires (desires which are her own in the minimal sense that they are states of her mind), but who also acts from these desires with a view to the value they have for her, given the things she cares about. Such an agent can *own* her desires in the further sense that she identifies with or endorses them, and she can also *disown* her desires inasmuch as she believes that acting on them would conflict with her identity and undermine the things she holds dear. By contrast, in the form of agency exemplified by the wanton, the agent's practical perspective just is (or is little more than) a succession of first-order desires. That is to say, her behaviors are "as much attributable to the particular desires that were, on the occasion, strongest, as they are to an agent in any important sense different from those desires" (Bratman 2007, p. 51).[53] There is, as it were, no *one* over and above these desires to which the behavior can be attributed, and nothing beyond those desires (and the occurrent stimuli which elicit them) is *in control* of how the agent ultimately responds to her present situation. To use a poignant phrasing of Nietzsche's, such an agent is "tethered by the short leash of its pleasures and displeasures to the stake of the moment" (Nietzsche 1874/2006, p. 125).

After all, Frankfurt's essay was not really about addiction. It was about the relationship he saw between free will, personhood, and a certain economy of desires: "When a *person* acts, the desire by which he is moved is either the will he wants or a will he wants to be without. When a *wanton* acts, it is neither" (Frankfurt 1971, p. 14). A person on Frankfurt's theory is thus any agent with a view or perspective on her own desires—an evaluative or normative outlook on her own agency with which she (the person) can be identified.

Recall that, on the DPA, addiction is a form of diminished personhood in that addiction undermines an individual's capacity for self-governance. Like Frankfurt's theory, the theory of personhood on which this conclusion is based defines personhood in terms of a capacity to occupy a "removed" perspective on one's own mental life and present situation. It is called the *Self-Governance Model of Personhood* (SGP).[54] But the SGP also differs from Frankfurt's

[53] Very much like Frankfurt's wanton, the first hypothetical creature Michael Bratman discusses in *Structures of Agency* (2007) is described as "pushed and pulled by its desires" and "an agent in only a minimal sense" (p. 51).

[54] The SGP emerges out of a distinction of Jenann Ismael's (2016) between two kinds of complex, physical systems—between *self-governing* systems and *merely self-organizing* systems—the finer details of which can be set aside for present purposes. Ismael's central claim is that, among physical systems which exhibit the kind of system-wide coordination that is suggestive of goal-directedness or intentionality, only in self-governing systems is that

theory in at least two ways. First, it emphasizes that, beyond merely occupying such a perspective, personhood also involves *governing* oneself in accordance with this self-reflective viewpoint. It is not enough that the agent occupy this perspective on her own agency; she must also use it to regulate how she thinks and behaves. Second, this capacity for self-governance extends well beyond a capacity to form second-order desires. Specifically, it involves (a) taking a *temporally extended perspective* on one's agency over time, (b) authoring a *personal narrative* and (c) negotiating and enforcing *personal rules*. According to the SGP, an individual is a person if, and to the extent that, she constructs, occupies, and utilizes this cognitively rich perspective "so as to maintain coherence among [her thoughts and behaviors] and keep them aligned with, or at least not subversive of, meaningful narratives told by the person to herself and, in fragments and allusions, to others" (Ross 2017, p. 2).

The role of personal narrative in the SGP is worth emphasizing. A personal narrative is quite literally the story an individual tells herself and others about who she is. This includes her values and priorities, who she would like to become, how she understands the world and her place within it, what she thinks is possible, permissible, required, and forbidden, and basically anything else that might find a place in an autobiographical account of a person's life. But besides being the source material for dinner parties and job interviews, the narrative also serves a central agential function: it sets evaluative and normative standards against which human beings can regulate their thoughts and behavior. For example, someone who sees herself as a good friend, or wishes to see herself as one, will sometimes be faced with situations that elicit "wayward" desires—desires, say, to stay home and watch a movie rather than help her friend with the arduous task of moving. It would no doubt be more enjoyable in the short term to stay home, but doing so would strain against the identity she recognizes, or wishes to recognize, as her own. She could save face by coming up with an excuse, but that will not make it any easier to see herself as a good friend. Ultimately she will need to negotiate with herself about how she is to respond to such circumstances, since that will in large part determine whether she is or is on her way to becoming a good friend. She will therefore need to work out certain principles, or

---

coordination due, at least partly, to *centralized* information processing and control. Contrary to the view that agency, selfhood, intentionality, and related concepts are completely *distributed* phenomena, Ismael maintains that in a self-governing system there is literally some *one* in charge—a centralized causal and information hub responsible for developing, maintaining, and enforcing a vision or plan for the system as a whole.

personal rules, which connect her moment-to-moment decisions to her long-term aspirations and identity.

In this way an individual's personal narrative brings with it a deliberative perspective that is richer than and, in a very important sense, "at a remove" from her present circumstances and first-order desires. It is because of her personal narrative, and all of the promises, aspirations, and duties articulated within it, that she is not tethered to the stake of the moment. By consulting this narrative she can appreciate what else at stake in her present choice—e.g., her status as a good friend. Jenann Ismael describes this advantage of narrative-driven self-governance in the following way:

> the choice-governed aspects of human behavior do not just depend on the immediate stimulus, but are open to influence from an in principle unlimited number of idiosyncratic sources in our personal past[, where] the bearing of this information on behavior is filtered through a quite complex set of higher-order principles for choices (goals, values, priorities, beliefs about who we are and who we want to be). (Ismael 2016, p. 95)

The personal narrative thus serves as an informationally rich layer of mediation between stimulus and response that effectively "decouples" the agent from the occurrent stimuli in her environment (ibid., p. 28), granting her a much greater degree of freedom over how she responds to those stimuli. To lend some precision to this idea, consider the contrast Ismael draws between "temporally thick" and "temporally thin" agency:

> Think of a temporally extended self as a population of individual temporally thin agents. And think of the collective goals as goals pertaining to the temporally thick agent (e.g., finishing a large project, being healthy, saving money), and the temporally thin agents as having more mundane, short-term interests (the crème brulee, the lazy Sunday, the expensive dinner out). Now consider the relationship between the utility calculation carried out by the deliberative perspectives of the individual temporally thin agents. It's in the interests of the long-term collective agent (that is constituted by my temporal parts) that I go to the gym today, practice piano, or put in a few hours on my book; that is, that I make some progress toward overarching ends. But only if these aren't a one-off deal, only if my other temporal parts are cooperating and similarly engaged. A single day on the book or at the gym doesn't promote the collective goals by itself. And if I have no way of ensuring or enforcing the collective interests on them, it makes no sense for me not to spend the day at the races. (Ismael 2016, p. 70)

The contrast is very much like the one Frankfurt presents between the wanton and the unwilling and willing addicts, except here the two types of agent are, as it were, contained within a single individual. Whereas an individual's "thin selves", corresponding to Frankfurt's wanton, have

only short-term interests guided by occurrent stimuli, her "thick self" has long-term interests corresponding to the goals, values, principles, and identity expressed in her personal narrative. Left to their own devices, an individual's temporally thin selves would have her respond to each situation in whichever way promises the greatest immediate payoff. But if she acts from a temporally extended perspective, then she does not respond merely on the basis of competing immediate rewards, but also and primarily on the basis of the value that choosing those rewards now will have for her in the long run, given her goals, values, identity, and aspirations.

To summarize, the SGP defines personhood in terms of self-governance, where self-governance involves three processes: (a) occupying a temporally extended perspective on one's own agency over time, (b) authoring a personal narrative, and (c) negotiating with oneself, on the basis of that narrative, about what personal rules one should adopt and enforce.[55]

It is worth considering for a moment how this process of self-governance can go wrong, because doing so will bring us back around to the DPA's central claim: that addiction is a form of diminished personhood. For one thing, it takes time to detect the patterns in one's life that are inimical to one's self-professed goals. Someone who wants to become a famous writer may nonetheless commit precious little time to reading and writing. There are a number of ways she can be startled out of this sort of wishful thinking, not least of which is the way other people, particularly those whose opinions matter to her, respond to her. It can also happen indirectly through her observations of her peers, whose recognition in similar areas forces her to consider why she has not made similar achievements. It can also take a more self-reflective form—some honest soul-searching, in which she forces herself to consider whether she really wants to do what it takes (or has what it takes) to achieve her goal, or whether she should instead update her personal narrative to better reflect the things she really values doing (or is capable of doing). Eventually, as a range of social or personal "checks" call on her to reevaluate her commitment, she may decide to adopt the rule, "From now on, I will write for three hours every morning." But this rule may be unrealistic. Maybe it fails to accommodate her other commitments, or maybe it is simply too drastic a change to effect immediately. How she responds to her failure to enforce such a rule will matter a great deal. She could renegotiate her other commitments, establish exceptions for specific days, or make a number of other adjustments. But she may instead

[55] For a discussion of the role that self-prediction plays in motivating this negotiation and enforcement of personal rules, see Ainslie (2020).

respond to her failure as evidence that she is not a writer—that, if she were really a writer, she wouldn't have such a hard time getting herself to write for three hours every morning. Maybe she heard that an author she admires starts each day this way, or that her colleague does, and, unaware of the many years it took the author to cultivate that regimen, or that her colleague is full of it, she prematurely concludes that she simply isn't cut out for the writing lifestyle. No doubt many possibilities for achievement have been abandoned in just this fashion.

Indeed, it is just this sort of response to failed self-governance that, according to the DPA, leads individuals with addiction to give up on the possibility of recovery. Recall the cycle represented in Figure 3. Addictive behavior leads to neuroadaptation in the individual's reward learning system. As a consequence, she will experience greater difficulty resisting desires to engage in the same behavior going forward. As such, if she tries to enforce a personal rule against that behavior, or one that allows it only under restrictive circumstances, she will be more likely to fail. She is thus more likely to engage in addictive behavior again, leading to further neuroadaptation, making it even less likely that she will successfully enforce those rules. But that is not all. In response to recurrent failures, she will eventually reach the conclusion that she is simply not the sort of person who can self-govern—that, for someone like her, there is little to no value in setting personal rules in the first place. Using Ismael's notions of temporally thick and thin agents, Ross expresses this idea as follows: "The addict who has temporarily lost all confidence in her ability to legislate effective personal rules has collapsed into identity with her community of temporally thin sub-agents" (Ross 2017, p. 12). On the DPA, the extent to which an individual is addicted is the extent to which, through a cycle of behavior, neuroadaptation, and self-interpretation, she forfeits the deliberative perspective and self-governing process that is essential to personhood.

Section 2 – Dependencies and Self-Governance

Before presenting my objection to the DPA and, by extension, the broader framework of addiction as a diminished capacity to self-govern, I want to briefly acknowledge those aspects of the DPA with which I agree. In many ways, I think it is on the right track. For one, I agree that addictive behavior results in the kind of neuroadaptation Ross emphasizes, making it more difficult for individuals to enforce personal rules against the behavior in question. I also agree that individuals who repeatedly fail to enforce such rules are likely to interpret those failures as

evidence of the kind of person they are and what they are capable of, thereby influencing the way they represent themselves in (and govern themselves by) their personal narratives. Indeed, I commend the DPA's ability to draw this important bridging connection between the neurobiological and the choice-theoretic dimensions of addiction. Finally, I am also sympathetic to the conception of personhood developed in the SGP, and I even agree that many individuals with addiction are, at least sometimes, diminished persons in this technical sense.

What I disagree with is the idea that addiction is best understood as a form of diminished personhood. Sometimes the reason an individual with addiction continues to engage in addictive behavior is not that she has given up on self-governance, but is rather that she has *successfully* governed herself according to a particular vision of who she is, what her values and responsibilities are, and what she is capable of. And as I'll address in my replies to objections, this is not to be confused with the claim that individuals with addiction, *despite* their addiction, sometimes successfully govern themselves according to personal rules. Rather, what I am saying is that sometimes addiction *itself* is an expression of self-governing personhood. This may seem hair-splitting to some, but it is actually quite crucial. For if addiction is sometimes explained by the *type* of personal rules an individual uses to regulate her behavior, given the *type of person* she takes herself to be, then it is simply not correct to identify addiction with a diminished capacity for self-governance.

My suggestion draws on an alternative approach that emphasizes the role that a person's identity, values, goals, and social life sometimes play in stabilizing patterns of addictive behavior. Hanna Pickard, a leading proponent of this approach, has argued that "we cannot explain many cases of addiction without recognizing the *value* of drugs to people, including people who are addicted", and that "a sense of self and social identity can be a central part of the explanation of addiction" (Pickard 2020, p. 2). If there is anything to this idea, then a theory of addiction as diminished personhood is partial at best. It may very well be a good explanation of some cases of addiction, or of one aspect of some cases, but it fails to appreciate the myriad ways in which addiction can *fit into* and *be an expression of* self-governing personhood.

In what follows I describe one general way in which this sometimes happens. When a person represents herself as *dependent* on things like drugs, alcohol, or gambling, this will prompt her to negotiate and enforce personal rules that *prescribe* addictive behavior, rather than proscribing it. These are beliefs, contained in her personal narrative, concerning the things she

requires in order to achieve her goals, manage the stressors and responsibilities in her life, uphold her values, participate in interpersonal and communal relationships, and so on. A personal narrative that represents oneself as dependent in this sense delimits the range of possibilities to which one believes oneself to be availed, with the effect that addictive behaviors continue to be pursued partly because of the enabling connection the individual sees between those behaviors and the things she values, or, correlatively, the disabling connection she sees between not engaging in those behaviors and those same sources of value.

To be as clear as possible at the outset, my claim is not that dependencies are relevant to an explanation of addiction in all cases, but is only that they are "central to explaining addiction in *some* addicts some of the time" (ibid.). Moreover, dependencies are just one relatively straightforward example of a phenomenon that is much broader, and which is as potentially variable as are the concrete details of a particular person's life. They thus represent one general example of how addiction can be an expression of self-governing personhood.

For the purposes of this paper, I call a "dependency" any belief about oneself according to which one is capable of *x* only if one does *y*. A dependency thus represents certain things as possible or feasible for the agent only under certain conditions. The general form of a dependency is "I can't *x* unless I *y*." Here are some examples:

x	*y*
fall sleep on airplanes	take Dramamine
eat solid foods	wear dentures
walk	use crutches
perform well in aerobic sports	use an inhaler

The significance of dependences for the present context is that they are self-beliefs which identify certain actions as the means to valued ends. Inasmuch as the agent values those ends (and is rational), she will be instrumentally motivated to negotiate and conform to principles, plans, and rules that take her dependencies into account. A corollary effect is that the agent will not be motivated to consider alternative rules so long as she continues to value the mediate end and continues to believe that the prescribed action is the only feasible means of achieving that end. Lastly, dependencies may be action-guiding even if they are not true.

For example, if I value performing well at aerobic sports and believe I can't do so unless I use an asthma inhaler, I have a clear reason to enforce a rule on my behavior that prescribes

inhaler use before aerobic sports. Moreover, it would be a failure of self-governance to play without using the inhaler. Suppose I sometimes fail in this way. Maybe I worry about how my peers will treat me for needing an inhaler, and so the immediate payoff of deviating from my rule is that I won't have to deal with their bullying. But I'll pay for that decision as soon as I jump into the pool, perhaps severely in the form of an asthma attack, or less severely in the form of diminished athletic performance. On the other hand, if I choose to use the inhaler, then that choice is made not on the basis of the immediate payoffs of present alternatives, but rather on the basis of an enabling connection that I see between two actions which form a temporally extended sequence.

To be sure, the inhaler example and the others listed in the table above are disanalogous to addictive behaviors in at least two respects. First, using an inhaler does not lead to the kind of neuroadaptations that using amphetamines or opiates leads to, and second, using an inhaler does not produce euphoric pleasure. But this does not cut against thinking that things like amphetamines or opiates can be interpreted by the individuals who use them as serving valuable ends which they believe would otherwise be impossible or unfeasible. For even if using an asthma inhaler did lead to similar neuroadaptations and did produce euphoria, so long as I still believe that I must use the inhaler in order to perform well at aerobic sports, I will still see instrumental value in using the inhaler. For an example closer to reality, just consider how little changes, vis-à-vis this instrumental value, if "take Dramamine" in the above table were replaced with "take Valium".

Or consider the aspiring writer from Section One. Suppose that, instead of reaching the conclusion that she is not really a writer, she begins to seek out motivational aids. She already drinks coffee—this she believes she needs even to *get up* in the morning—but she also knows that a fair number of her academic peers smoke cigarettes while they write, and one morning she asks if she could take a drag. The feeling is like a match made in heaven. Some people experience adverse reactions to their first cigarette (e.g., nausea), but with just this first drag, our aspiring writer feels a measure of focus and motivation that she has never felt before. That morning, she continues to smoke as she writes what later becomes her first A+ assignment. From then on, smoking is part and parcel of her writing process. Sure, she also begins to smoke at times when she is not writing, but whenever she writes, she smokes, and whenever she plans to write, she plans to smoke.

It is also worth mentioning that the euphoric pleasure produced by addictive drugs and alcohol need not be ancillary to their instrumental value, particularly in circumstances in which that euphoric state is itself what enables the individual to do other things she values. Consider the following example—which, for what it is worth, is adapted from a real example of someone in my life. Suppose Kaitlin works long, grueling shifts in a high-intensity kitchen, where the temperatures sometimes exceed sauna levels and the demandingness of her supervisors knows no bounds. Co-workers either are unreliable or, if they are reliable, quit within a matter of weeks. Kaitlin prides herself on her dedication to her work, as she believes that her perseverance is among her most valuable traits. As such, and despite the behest of loved ones, she has little interest in seeking out alternative forms of employment. Besides, the pressure to pay off compounding debts she incurred earlier in her life make it so that the mere thought of risking stable employment for a potentially cushier job is terrifying. Not only does standing for long hours have negative effects on her physical well-being, causing persistent back and foot pain, but the sheer magnitude of her responsibilities also has effects on her mental health. She easily does the work of two or even three of her co-workers, and her willingness and ability to do so have only led to more responsibilities being piled on. Sometimes she stands up for herself when she feels that she is being especially overworked, but it is usually not worth the effort or the risk. Now here's the rub. Kaitlin believes that she would be incapable of enduring the chronic pain and stress of her work if she didn't drink, and drink fairly heavily, each and every evening. Sure, she enjoys drinking, and these euphoric and pain-reducing effects are the main reasons she continues to do so. But even so, her decision to drink involves more than a mere comparison of immediate payoffs between drinking and non-drinking alternatives. Kaitlin does not opt to drink for the simple reason that she'll feel better now. She also drinks because, if she didn't, she has no idea how she could keep this up. "It's all I have," she says. And if friends or family press her on the issue, she promptly (and fiercely) reminds them of the life she leads, day in and day out, to keep the bills paid—to pay off the debt while she still can, before it becomes a problem for her children. To be clear, Kaitlin believes that it would be best, hypothetically, if she weren't dependent upon alcohol in this way, due to the other negative effects it has, e.g., on her long-term health. But this point is moot so long as she genuinely believes that managing the psychological stress of her work life would be impossible if she did not drink, or too unfeasible to risk. Given the actual circumstances of her life, she believes that her best option is to keep

drinking, because at least that way she can feed, clothe, and insure her family, as well as protect them from her debts.

Kaitlin's rationale for her drinking bears clear similarities to the "self-medication hypothesis" of addiction (Khantzian 1985; 1997). According to that hypothesis, drugs and alcohol "offer a way of coping with intense negative emotions" and "severe psychological distress" (Pickard 2012, p. 41). It thus straightforwardly suggests that some people engage in addictive behavior purposively, i.e., because of the instrumental value it has for them. However, the self-medication hypothesis is controversial, and I want to quickly address that controversy in order to clarify what I am, and am not, endorsing.

Anna Lembke (2012) has argued, for example, that the self-medication hypothesis should be abandoned, at least in the case of individuals with comorbid mental disorders, because of the influence it has on the way these patients are treated, as well as its reinforcement of patients' "self-medication perspective" (p. 527). For one thing, she emphasizes empirical findings which show that substance use does not actually result in "sustained improvement in mood, cognition, or function", but does result in "greater non-compliance with treatment, and overall poorer outcomes" (ibid., p. 527). Lembke also argues that the hypothesis "encourages clinicians to target only the 'underlying' psychiatric illness and ignore addiction, on the grounds that once the depression or mania or anxiety or whatever it is, is treated, the substance use problems will resolve", and that it "likewise encourages patients to interpret all psychological suffering […] as originating from the underlying psychiatric disorder rather than from dependence on or withdrawal from addictive substances" (ibid., pp. 526-27). Acknowledging that providers can more easily cultivate a "therapeutic alliance" with patients by endorsing the self-medication hypothesis, Lembke urges that "the risks of colluding in a narrative which is poorly founded outweigh the risk of making our patients angry" (ibid., p. 527).

Lembke's objections are thus primarily concerned with the clinical upshot of the self-medication hypothesis and with the claim that substance use serves a therapeutic function. In response, I want to draw attention to two things. First, for the purposes of this paper, I aim to remain agnostic as to whether substance use ever serves a therapeutic function. What I do endorse is that some individuals with addiction *believe* that this is true. Indeed, Lembke agrees that many individuals with addiction believe that their drug or alcohol consumption has some such instrumental value. She says, "the majority of psychiatric patients retrospectively endorse

[the association of] relief of psychiatric symptoms with substance use". Moreover, the worry she raises in connection with treatment is precisely that providers should not lend credibility to "patients' self-medication perspective" (ibid., p. 527). The concern that providers should not be "colluding" in an ill-founded narrative can be a concern only if that narrative is adopted by at least some patients.

Second, in this paper I am concerned only with the descriptive account of what addiction is, and so I do not mean to advocate for any particular treatment method. Lembke argues that treatment professionals should not validate their patients' belief that substance use serves a therapeutic function and that addiction is best understood as a symptom of the patient's underlying psychiatric disorder (or other source of chronic psychological distress). But I briefly want to flag a second way in which treatment professionals might endorse the self-medication hypothesis: they may simply believe *that the patient believes* that her substance use serves a therapeutic function, and use this information to tailor treatment to that individual's particular epistemic situation.[56] While this matter falls outside of the scope of the present paper, I believe it is only this latter kind of endorsement that my account may be taken to support.

Having said that, I want to clarify that the operative notion of dependency in this paper is not restricted to specifically therapeutic functions. As the aspiring writer example already suggests, addictive behavior may instead be chosen as the means to achieving professional aspirations and, more broadly, cultivating and maintaining the individual's personal narrative. Of particular relevance here is Hanna Pickard's recent paper, "Addiction and the Self" (2020), in which she argues that individuals sometimes choose to engage in addictive behavior precisely because they self-identify as addicts. For one, addictive behavior and the broader lifestyle of addiction can be a source of social reward and community. Although theorists typically emphasize the negative impacts of addiction on social life, Pickard argues that it can also create and be a focus of activity for meaningful and rewarding relationships. One example is the comradery of smoking huddles and bars, but most poignant is Pickard's description of "addicts who are members of some of the most vulnerable and marginalised drug user communities, such as long-term homeless poly-drug heroin users":

[56] One obvious way a provider might do this is by "educating patients about the limits of the [self-medication hypothesis]", just as Lembke suggests (ibid., p. 527).

> [M]any such addicts have long lost most if not all relationships with individuals and communities who are not drug-using. Their user community may be all that they have left; and the relationships between members of such communities can be deep and meaningful. Living on the margins of society, addicts may love, protect, and care for each other, while they face their collective daily need for drugs in a context of poverty, homelessness, disease, disability, and police harassment and violence. Their identity as addicts is precisely what binds them together: quitting using would involve quitting the community and these relationships. Meanwhile the lack of treatment, housing, health care, and employment opportunities, combined with the intense stigma surrounding this form of addiction, means that in reality, abstinence does not guarantee a better life, let alone the possibility of replacement relationships of comparable commitment and meaning. Hence, for addicts such as these, their addiction is a way of belonging to a group and maintaining social bonds, when they would otherwise face the prospect of severe social isolation and loneliness. […] Self-identifying as an addict can therefore have value because it brings with it a community of people who care. (Pickard 2020, pp. 11-12)

In such a case, addictive behavior is intimately connected with an individual's social life and functions as a source of the kind of social reward that comes from membership and participation in a community. And the dismal prospects of finding similar bonds elsewhere lends further credence to the belief that continued drug use is the only feasible means to rewarding and meaning relationships. Here the motivation to engage in addictive behavior would therefore fall within the broader contours of my notion of a dependency, inasmuch as it has *instrumental* value as a means to social reward, and would thus be an instance in which addictive behavior is an expression of self-governing personhood.

Drawing on Ian Hacking (1996), Pickard also argues that self-identifying as an addict may be tantamount to identifying as a *social kind.* Social kinds, including races, genders, nationalities, and political affiliations, to name just a few, are ways of categorizing individuals (including oneself) that tend to be "associated with specific sets of believes, values, and behaviors, to which members are expected to conform *in virtue of their membership"* (ibid., p. 10). These expectations, or *group-specific norms,* prescribe "what group members *are supposed to be like*" (ibid., p. 10). She argues, "if a person *self-identifies* as a member of a social kind they are likely to (explicitly and implicitly) *self-regulate* to ensure their beliefs, values and behaviour conform to the group-specific norms in question" (ibid., p. 10). As such, Pickard argues, some individuals with addiction may engage in addictive behavior because that is "what addicts are supposed to do" (ibid., p. 11).

Pickard's suggestion that identification with a social group can lead individuals to adopt and enforce group-specific norms (both on themselves and on other group members) is also backed by a growing body of interdisciplinary research that collectively suggests that humans are equipped with a psychological capacity to detect and internalize norms in their social environments, thereby becoming strongly motivated to comply with and enforce compliance with those norms (Kelly & Setman 2020; Sripada & Stich 2007). The big question is whether self-identification as an addict results in norm-internalization in this sense. But if it does, then it is a small step to say that individuals who have internalized these norms would self-govern in ways corresponding to their prescriptions.[57]

Recall that on the DPA individuals with addiction come to view themselves as incapable of self-governance after repeatedly failing to enforce personal rules. Another possibility, suggested by the notion of a dependency, is that an individual may come to view herself as incapable of valued ends—e.g., achieving goals, managing stress, fulfilling responsibilities, maintaining social identities—without engaging in addictive behavior. Indeed, one way an individual may reach this conclusion is by observing that she has repeatedly failed to do these things when she abstains from that behavior. The reasons for interpreting her behavior in this way could even include the very difficulty of enforcing personal rules that the DPA emphasizes. That is to say, the very same struggle to legislate and enforce effective rules against addictive behavior may lead some individuals to represent themselves as incapable of doing the things they value while simultaneously struggling to enforce such rules. It may not be worth the effort or the risk it poses to her other commitments. Another related possibility is that the distracting and attention-consuming nature of somatic cravings becomes an impediment to doing just about anything until the cravings subside. Clearly, one solution to this problem is to satisfy the craving.

[57] Whether it would also be an instance of a dependency is less clear, however. As I am using the term, a dependency is a belief in virtue of which individuals become *instrumentally* motivated to behave in ways they believe will enable them to achieve desired ends. But the motivation that results from norm-internalization may not be instrumental. Although some researchers argue that compliance with norms is entirely motivated by the threat of social punishment and the promise of social reward (Bicchieri, Muldoon, & Sontuoso 2018), others have argued that, once norms have been internalized, they motivate individuals *intrinsically* (Kelly & Davis 2018). On such a view, "People will be motivated to comply with and enforce a such rule for its own sake, and experience an impetus to do so that is independent of external circumstances or the perceived likelihood that they will receive social sanctions even if they flout the norm" (Kelly & Setman 2020).

For an individual who does not have the money or time to dedicate themselves to the clinical process of recovery, this may be the only feasible option. Inasmuch as those cravings are preventing her from engaging in and attending to the things she values, the instrumental value of addictive behavior is straightforward.

But as the examples above have suggested, dependencies may also arise from a more direct connection between addictive behavior and desired ends. The individual may believe that it is simply the best way to keep on trucking, as in Kaitlin's case, or that it is what she needs to achieve the required level of commitment to her aspirations, as with the aspiring writer. It may also, as Pickard argues, be the only feasible means to meaningful and rewarding relationships, or it may simply be "what addicts are supposed to do". In all of these cases, the individual engages in the very sort of narrative-driven self-governance that the DPA takes to be essential to personhood: she sees addictive behavior as enabling her (a) to be who she is and (b) to achieve the things she most deeply values, given who she is. This view of oneself as dependent is thus a piece of the personal narrative in terms of which one exercises self-governance through the implementation of personal rules. A personal rule that prescribes drug use on the basis of a dependency would not be a forfeiture of self-governance, but an instance of it.

Section 3 – Objections and Replies

I want to close by anticipating and replying to a few objections.

1) Inasmuch as addictive behavior causes neuroadaptations that make it more difficult to enforce personal rules—indeed, inasmuch as this is what makes those behaviors addictive—shouldn't this phenomenon take center stage in an account of addiction, as it does on the DPA?

I want to be as clear as possible that I do think addictive behavior has lasting neurobiological effects. Still, I think that the identification of addiction with diminished personhood does not follow from that fact. A person may struggle to enforce personal rules in this or that domain, without being led to think of herself as incapable of self-governance. Indeed, she may respond by self-governing in ways that take this struggle into account: she may give up on thinking that she can lead the life she wants without using a specific drug, and so embrace this fact about herself

in something like the way near-sighted people embrace the fact that they cannot drive without glasses. Her inability to enforce rules against using the drug thus comes to be interpreted in a manner analogous to a disability.

So, I am happy to agree with the claim that part of addiction, and part of what makes certain behaviors addictive, is that individuals with addiction would experience greater difficulty enforcing personal rules against those behaviors, should they choose to do so. My claim has been that there is more to addiction than this. I think that the identification of addiction with diminished self-governance may reflect a bias in the literature in favor of *recovering* addicts—individuals who, having rejected their addiction, struggle to change their behavior accordingly. In other words, the literature tends to focus on the explanatory and practical challenges presented by "unwilling addicts" and tends to have relatively little to say about "willing addicts". The fact that people do sometimes embrace their addictions strongly suggests that we have not reached the end of the road in our explanation of addiction once we have a good account of the way it undermines self-governance.

2) Aren't these "dependencies" just ad hoc rationalizations (a la Haidt 2001)? I'm treating them like they play a robust causal role in decision-making, whereas it seems more likely that they are just what a person tells herself to rationalize her failures to self-govern.

For one thing, this objection cuts both ways, since personal narrative is also supposed to play a robust causal role on the SGP. But even if dependencies are not an efficient cause of addictive behavior, or something an individual explicitly considers when she engages in those behaviors, they can still play a role in the closing off opportunities for critical self-reflection which might otherwise arise. Relatedly, personal narratives and the rules they give rise to do not just serve important functions for internal self-regulation. They also serve important functions for interpersonal self-regulation. These are stories we tell not only to ourselves, but also to others—others who are thus in a position to recognize and validate our story, or call us out on it. Other people, particularly those whose normative authority we recognize, have the ability to hold us to higher standards than we hold ourselves to. As it is sometimes put, human agency is shaped and scaffolded by our social practices (McGeer 2015, Vargas 2013, Zawidzki 2013).

3) What I'm describing isn't "true addiction", it is a stage *prior* to true addiction—a time when the person is not yet really addicted, but is engaging in behavior that will result in addiction over time. As Ross says (2020), "[t]hat addicts typically choose the behavior that triggers their addiction is disputed only by skeptics about free will in general" (p. 5).

Surely there is a transition into addiction, but I for one doubt that there is any clear and distinct point of transition. Addiction comes in degrees, and the extent to which a person is addicted is not just the extent to which she struggles (or would struggle) to enforce personal rules against addictive behavior. That is part of what stabilizes patterns of addictive behavior, part of what explains the persistence of that behavior despite significant negative consequences, but it is far from the only thing. It can be tempting to think of all cases of addiction as teleologically oriented toward a struggle of self-governance. But even if this is commonly what happens, it should not set the bar for "true addiction".

4) What I am describing is the "discovery of effective personal rules" (Ross 2020, p. 6), i.e., a person who has regained self-governance and is no longer addicted.

Like the previous objection, this one turns on the question of where we are to draw the line between "true addiction" and something else, like "controlled use"—i.e., the use of drugs, alcohol, etc., in a way that avoids the most severe negative consequences and is more-or-less sustainable. The fact that the negative consequences, in terms of which addiction is standardly defined, are therefore largely abated, could be taken to suggest that this would no longer be a case of addiction. My own intuition is that, for an individual like Kaitlin from Section Two, the fact that her drinking has been managed so as to be consistent with and maybe even in service of her goals and commitments does not mean she is not addicted to alcohol. For one, there are still negative consequences to her drinking, even if they are outweighed and mitigated, and she would still experience difficulty changing her drinking patterns if she decided that there were better ways to manage her pain and stress. Also, saying that such cases are not "true addictions", because the negative consequences are outweighed, begs the question against theories of addiction which argue that addiction is sometimes explained by the value of drugs. The reason such a theory is incorrect cannot be that, if someone is truly addicted, the negative consequences

would outweigh the positives, since part of what is at issue is precisely whether addiction can sometimes be the subjectively rational option.

5) Inasmuch as a person *falsely* believes themselves to be dependent upon something in order to achieve the things they value, and enforces personal rules which reflect this falsehood, isn't their capacity for self-governance still diminished in some sense?

I do think there is a sense in which such a person's agency is diminished if she is governs herself according to false beliefs, but the problem is not that her capacity to negotiate and enforce personal rules is diminished. A person who believes that she *cannot* do something and devises personal rules that take this incapacity into account is a perfectly good self-governor, though she may be diminished in epistemic domains. Someone who is ignorant, misled, or hallucinating is not for that reason less capable of regulating her behavior according to personal rules.

CONCLUSION

In this dissertation I have argued for a number of theses related to responsibility, norm psychology, and personhood. Although most of the papers argue for "standalone" theses, in the sense that their truth does not depend the truth of the others, the five papers collectively illustrate a broader view of humans as (a) responsible agents who are (b) self-governing and (c) equipped with a capacity for norms, and whose agency (d) centers on dynamic responsiveness to corrective feedback. Drawing on this broader picture, I have tried to shed light on ethical questions about our social practices and technologies, as well as descriptive questions about the nature of substance use disorder.

Most centrally, my project has been an attempt to show that forward-looking considerations are relevant for responsibility, not merely because certain consequences of our responsibility practices are desirable, but primarily because of a connection which I have argued exists between relationships, norms, and learning. One of Strawson's key insights in "Freedom and Resentment" was that, in order for an agent to be justifiably held responsible as a participant in norm-governed relationships, the agent must be an "appropriate object" of a basic demand for good will. Importantly, this status requires that the agent possess a capacity to understand and live up to that basic demand, where the demand may take on a variety of forms depending on the specific relationship. These would be the specific norms governing that relationship. In paper one I argued that the capacity in question is an "intelligent" capacity in Ryle's sense, which is to say that it is not merely a first-order disposition to recognize and respond to certain reasons, but is also a second-order capacity to refine and adjust that first-order disposition in response to corrective feedback. In my view, a paradigmatic form this feedback takes is the social feedback humans receive through normative responses like praise and blame. Similarly, in paper two I argue that humans can be justifiably held responsible even for attitudes and actions which result from automatic, unconscious, and intuitive aspects of human psychology, so long these intuitive processes can thereby be attuned to reasons over time.

A closely related goal has been to show why theories like the answerability theory and deep self theories—what I call "self-direction" theories—are mistaken about the role that control plays in responsible agency. Self-direction theories present a compelling and elegant alternative to control-based theories, including forward-looking theories and the broader reasons-

responsiveness approach, and in many ways my project has taken shape in response to this challenge. In paper three I argued that deep self theories are really variants on the broader reasons-responsiveness approach, which is to say that control is already at issue in notions like self-expression. And in paper one I argued that, in order for an agent to be eligible for normative responses like praise and blame, the agent must be capable of having substantive responsibilities—in Strawson's phrasing, they must be "appropriate objects" of the basic demand for good will. Self-direction theories, although they do not deny this requirement, overlook the constraint it places on a satisfactory account of responsible agency. On their view, an agent is a responsible agent just in case she is capable of acting in ways that reflect her judgment or express her self, because this allows the agent's thought and behavior to be attributed to her in the way required for normative appraisal. I have argued that this is a mistake, because normative responses evaluate agents as having met, exceeded, or violated norms *which the agent is legitimately subject to.* A complete account of responsible agency, then, must include those features of responsible agents which qualify them for interpersonal normative expectations, demands, and standards. I have argued that a merely self-directed agent would fail to meet this requirement if she is incapable of recognizing and correcting her own normative failings, even if she can still be evaluated against normative standards in a technical sense.

On the resulting picture, an agent is a responsible agent only if she can control her thought and behavior over time, so as to regulate herself according to norms of which she presently falls short. One way of putting this idea is that, while synchronic leeway freedom—the capacity at a time to perform alternative actions—is not necessary for responsibility, *diachronic* leeway freedom is necessary, because without it agents could not reasonably be expected to behave differently than their standing first-order dispositions allow. This strikes me as an essential feature of our eligibility for normative responses, inasmuch as these responses are ways of holding agents accountable to norms even and especially when they violate them.

Importantly, my claim is not that our responsibility practices are justified insofar as they tend to have desirable consequences, but is rather that they are justified by certain capacities of responsible agents which make them eligible participants in norm-governed relationships. In my view, these capacities depend in turn on the agent's exposure to a system of social practices—what Manuel Vargas calls the agent's "moral ecology". Unlike Vargas, however, I do not think that these practices are justified by the desirability of cultivating responsible agency, be it at the

individual or collective level. Rather, they are justified by the norms which agents are legitimately subject to—from putatively universal moral norms to context-, role-, and relationship-specific norms—where the legitimacy of those norms partly depends on capacities that our social practices cultivate and maintain. Legitimate norms, then, do the basic justificatory work, but their legitimacy is sensitive to the actual capacities of agents—i.e., whether the agents are capable of thinking and acting as they putatively should. Circling back to the quotation from Dennett in the introduction, our status as responsible, norm-governed agents is socially contingent, and it is contingent, specifically, on our continued willingness to treat one another as agents who are "up to the task" of interpersonal norms. Whether we are capable of fulfilling the expectations we have of one another, and so too whether those expectations are legitimate, depends on whether we continue treating one another as being so capable.

Finally, my project has also taken up issues related to emerging social technology and to the role of personhood in addiction. While papers four and five to not speak to the basic question of what makes humans responsible, they do contribute to the broader account of human agency in which that basic question is situated. So, in paper four, I argue that human norm psychology has implications for how we should relate to "social robots"—artificial agents designed to participate in relationships with humans. I argue that, like humans, social robots should be equipped with a capacity to recognize and respond to normative feedback. This is arguably a high bar, but in the event that social robots are successfully designed to simulate co-participation in social relationships, humans will tend to attribute attitudes to social robots and to care about the attitudes they thus believe are being directed toward them. Part of human sociality is our willingness to enter into relationships, even and often because of the way these relationships require us to be vulnerable with one another. But this willingness is conditioned on our belief that the persons we so relate to will be responsive to us in ways that make it possible to safely navigate this vulnerability. I argue that the same condition should apply in human-AI relationships. Importantly, this responsiveness involves a second-order capacity to update the way we think and behave so as to better meet others' expectations of us, especially considering that many of these expectations are subject to change. While it is not an issue I take up directly in this dissertation, a future direction for my research program is to consider in more detail this aspect of interpersonal relationships—that they are, so to speak, "normatively dynamic"—and to see how this might contribute to a fuller understanding of responsible agency.

The fifth and final paper of the dissertation argued that addiction is sometimes an expression of self-governing personhood, and therefore is not a form of diminished personhood. The paper resists a common narrative about addiction as being a form of akrasia in which the agent acts against her own better judgment. Addiction is thus construed as a diminished capacity for self-governance or self-control. While this is certainly a central aspect of many cases of addiction, it fails to appreciate the ways in which addiction interacts with a person's identity and goals, especially in cases where the agent believes that the things she values would not be feasible if she did not continue to engage in addictive behavior. What I conclude is neutral with respect to the question of responsibility for addictive behavior, but it does suggest that in some cases of addiction, or at some stages of addiction, agents are choosing in ways that reflect their judgment about what is possible and what is valuable. On my view, as the earlier papers suggest, this alone would not be sufficient for responsibility. But it does at least open the door to considering whether normative responses may sometimes be appropriate, perhaps especially in close interpersonal relationships with individuals who have addictions, but only if it should turn out that our relationship with such a person bears the kind of motivational efficacy that could encourage change. To be clear, I believe there is reason to doubt that anything is wrong with addictive behaviors *per se.* But it does at least seem reasonable to suggest that individuals with addiction are capable of participating in certain relationships, such as friendships or romantic partnerships, where legitimate expectations may arise between the parties about whether continuing to engage in addictive behavior is permissible, particularly if the behavior impairs the relationship. This is, no doubt, an ethically fraught and controversial issue. Indeed, as a brief autobiographical note, it is the one which first led me to study responsibility, partly because of the pressing philosophical challenge it raises about how we should understand human agency and our social practices in the face of behavioral disorders like addiction, but also because of the pressing social challenge it raises about how we should understand our actual relationships with such persons—how to talk with them and how to care for them within the bounds of what is reasonable to expect.

In sum, the dissertation has tried to illuminate features of human agency which I believe make us capable of navigating, and are implied by, the rich social ecology we presently inhabit. I have described us as having a form of agency that goes beyond mere authorship and self-expression, as agents who are capable of norms, who are self-governing persons, and who are

responsive to social corrective feedback which functions to cue us into opportunities for learning that go well beyond instinct and ordinary adaptation. More importantly, I have argued that these are among the features which make us capable of participating in norm-governed relationships and therefore eligible for normative responses, and I have applied this broader picture to address issues AI ethics and philosophical psychology.

REFERENCES

Ainslie, G. (2020). Willpower With and Without Effort. *Behavioral and Brain Sciences*, *44,* e30-81. https://doi.org/10.1017/S0140525X20000357

Akalin, N., Kristoffersson, A., & Loutfi, A. (2019). Evaluating the Sense of Safety and Security in Human-Robot Interactions with Older People. In O. Korn (Ed.) *Social Robots: Technological, Society and Ethical Aspects of Human-Robot Interaction.* Switzerland: Springer Nature Switzerland AG.

Behdadi, D., & Munthe, C. (2020). A Normative Approach to Artificial Moral Agency. *Minds and Machines, 30,* 195-218. https://doi.org/10.1007/s11023-020-09525-8

Beller, S. (2010). Deontic Reasoning Reviewed: Psychological Questions, Empirical Findings, and Current Theories. *Cognitive Processing*, *11,* 123-132. https://doi.org/10.1007/s10339-009-0265-z

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. *Science Robotics, 3*(21). https://doi.org/10.1126/scirobotics.aat5954

Berto, F., & Jago, M. (2018). Impossible Worlds. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), URL = <https://plato.stanford.edu/archives/fall2018/entries/impossible-worlds/>.

Bicchieri, C., Muldoon, R., & Sontuoso, A. (2018). Social Norms. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy* (Winter 2018 edition), URL = <https://plato.stanford.edu/archives/win2018/entries/social-norms/>.

Boyd, R. (2018). *A Different Kind of Animal: How Culture Transformed Our Species.* Princeton: Princeton University Press.

Boyd, R., & Richerson, P. J. (2005a). Solving the Puzzle of Human Cooperation. In S. C. Levinson & P. Jaisson (Eds.) *Evolution and Culture*. Cambridge, MA: MIT Press, 105–132.

Boyd, R., & Richerson, P. J. (2005b). *The Origin and Evolution of Cultures*. New York: Oxford University Press.

Boyd, R., & Richerson, P. J. (1992). Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups. *Ethology and Sociobiology*, *13*(3), 171-195. https://doi.org/10.1016/0162-3095(92)90032-y

Bratman, M. (2007) *Structures of Agency: Essays.* New York: Oxford University Press.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Chudek, M. & Henrich, J. (2011). Culture–Gene Coevolution, Norm-Psychology and the Emergence of Human Prosociality. *Trends in Cognitive Sciences*, *15*(5), 218-226. https://doi.org/10.1016/j.tics.2011.03.003

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

Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology*, *12*(3), 235-241. https://doi.org/10.1007/s10676-010-9221-y

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

Cummins, D. D. (1996). Evidence for the Innateness of Deontic Reasoning. *Mind & Language*, *11*(2), 160-190. https://doi.org/10.1111/j.1468-0017.1996.tb00039.x

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

de Graaf, M. M. A., Ben Allouch, S., & Klamer, T. (2015). Sharing a life with Harvey: Exploring the acceptance of and relationship building with a social robot. *Computers in Human Behavior*, *43,* 1-14. https://doi.org/10.1016/j.chb.2014.10.030

Dennett, D. (2003). *Freedom Evolves.* NY: Penguin Group.

Dennett, D. (1981). *Brainstorms.* Cambridge: MIT Press.

Dereshev, D., & Kirk, D. (2017). Form, Function and Etiquette – Potential Users' Perspectives on Social Domestic Robots. *Multimodal Technol. Interact., 1*(2), 12. https://doi.org/10.3390/mti1020012

Duffy, B., Rooney, C., O'Hare, G., & O'Donoghue, R. (1999). What is a Social Robot? *10th Irish Conference on Artificial Intelligence & Cognitive Science*. University College Cork, Ireland.

Edmonds, B. (1998). Modeling Socially Intelligent Agents. *Applied Artificial Intelligence,12*(7-8): 677-699. https://doi.org/10.1080/088395198117587

Edmonds, B. (1997). Modelling Socially Intelligent Agents in Organisations. *The AAAI Fall Symposium on Socially Intelligent Agents*. Cambridge, MA.

Fehr, E., & Falk, A. (2002). Psychological Foundations of Incentives. *European Economic Review, 46*(4-5), 687-724. https://doi.org/10.1016/S0014-2921(01)00208-2

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://doi.org/10.2307/2023833

Frankfurt, H. (1971) Freedom of the Will and the Concept of a Person. *The Journal of Philosophy, 68*(1), 5-20. https://doi.org/10.2307/2024717

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

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (Eds.) (2005), *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*. Cambridge, MA: MIT Press.

Greene, J. (2017). The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition, 167*, 66-77. https://doi.org/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.* Cambridge: MIT Press.

Hacking, I. (1996). The looping effects of human kinds. In D. Sperber, D. Premack, & A. James Premack (Eds.), *Causal Cognition: A Multidisciplinary Debate*, 351–395. Oxford: Oxford University 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://doi.org/10.1037//0033-295X.108.4.814

Haidt, J., Björklund, F., Murphy, S. (2000), Moral dumbfounding: when intuition finds no reason. *Lund psychological reports, 1*(2). Department of Psychology, Lund University.

Henrich, N., & Henrich, J. (2007). *Why Humans Cooperate: A Cultural and Evolutionary Explanation*. Oxford: Oxford University Press.

Heyes, C. M. (2018). *Cognitive Gadgets: The Cultural Evolution of Thinking*. Cambridge, MA: Harvard University Press.

Heyman, G. M. (2009). *Addiction: A Disorder of Choice*. Cambridge, MA: Harvard University Press.

Hieronymi, P. (2020). *Freedom, Resentment, and the Metaphysics of Morals.* NJ: Princeton University Press.

Höflich, J., & El Bayed, A. (2015). Perception, Acceptance, and the Social Construction of Robots—Exploratory Studies. In J. Vincent, S. Taipale, B. Sapio, G. Lugano, and L. Fortunati (Eds.), *Social Robots from a Human Perspective.* Switzerland: Springer International Publishing AG.

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,* 106-133. Oxford: Oxford University Press.

Hyman, S. E. (2005). Addiction: A disease of learning and memory. *The American Journal of Psychiatry, 162*(8), 1414–1422. https://doi.org/10.1176/appi.ajp.162.8.1414

Ismael, J. (2016). *How Physics Makes Us Free.* Oxford: Oxford University Press.

Josephs, M., Kushnir, T., Gräfenhain, M., & Rakoczy, H. (2016). Children Protest Moral and Conventional Violations More When They Believe Actions Are Freely Chosen. *Journal of Experimental Child Psychology, 141,* 247-255. https://doi.org/10.1016/j.jecp.2015.08.002

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*(9): 697-720. https://doi.org/10.1037/0003-066X.58.9.697

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

Khantzian, E. J. (1997). The self-medication hypothesis of substance use disorders: A reconsideration and recent applications. *Harvard Review of Psychiatry, 4*(5): 231–244. https://doi.org/10.3109/10673229709030550

Khantzian, E. J. (1985). The self-medication hypothesis of addictive disorders: Focus on heroin and cocaine dependence. *The American Journal of Psychiatry, 142*(11), 1259–1264. https://doi.org/10.1176/ajp.142.11.1259

Karami, A. B., Sehaba, K., & Encelle, B. (2016). Adaptive Artificial Companions Learning from Users' Feedback. *Adaptive Behavior*, *24*(2), 69-86. https://doi.org/10.1177/1059712316634062

Katz, J. E., Halpern, D., & Crocker, E. T. (2015). In the Company of Robots: Views of Acceptability of Robots in Social Settings. In J. Vincent, S. Taipale, B. Sapio, G. Lugano, & L. Fortunati (Eds.), *Social Robots from a Human Perspective.* Switzerland: Spring International Publishing.

Kelly, D. (forthcoming). Two Ways to Adopt a Norm: The (Moral?) Psychology of Avowal and Internalization. In M. Vargas and J. Doris (Eds.), *The Oxford Handbook of Moral Psychology*. Oxford: Oxford University Press.

Kelly, D. (2020). Internalized Norms and Intrinsic Motivation: Are Normative Motivations Psychologically Primitive? In C. Todd & E. A. Wall (Eds.), *Emotion Researcher,* 36-45. International Society for Research on Emotion.

Kelly, D. & Davis, T. (2018). Social Norms and Human Normative Psychology, *Social Philosophy and Policy*, *35*(1), 54-76. https://doi.org/10.1017/S0265052518000122

Kelly, D., and Setman, S. (2020). The Psychology of Normative Cognition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020 edition), URL = <https://plato.stanford.edu/entries/psychology-normative-cognition/>

Lembke, A. (2012). Time to Abandon the Self-Medication Hypothesis in Patients with Psychiatric Disorders. *The American Journal of Drug and Alcohol Use, 38*(6), 524-529. https://doi.org/10.3109/00952990.2012.694532

Leshner, A. I. (1997). Addiction is a brain disease, and it matters. *Science, 278*(5335): 45–47. https://doi.org/10.1126/science.278.5335.45

Levy, N. (2007). The Responsibility of the Psychopath Revisited. *Philosophy, Psychiatry, and Psychology*, *14*(2), 129–38. https://doi.org/10.1353/ppp.0.0003

Lycan, W. G. (1990). The Continuity of Levels of Nature. In W. G. Lycan (Ed.), *Mind and Cognition: A Reader*. Oxford: Blackwell Publishers, 77–96.

Maartje M. A. de Graaf, Somaya Ben Allouch & Jan A. G. M. van Dijk (2019). Why Would I Use This in My Home? A Model of Domestic Social Robot Acceptance. *Human–Computer Interaction, 34*(2), 115-173. https://doi.org/10.1080/07370024.2017.1312406

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MA: The MIT Press.

McAdams, D. P. (2019). 'First we invented stories, then they changed us': The Evolution of Narrative Identity. *Evolutionary Studies in Imaginative Culture, 3*(1), 1-18. https://doi.org/10.26613/esic.3.1.110

McGeer, V. (2018). Intelligent Capacities, *Proceedings of the Aristotelian Society, 118*(3), 347-376. https://doi.org/10.1093/arisoc/aoy017

McGeer, V. (2015). Mind-Making Practices: The Social Infrastructure of Self-Knowing Agency and Responsibility, *Philosophical Explorations, 18*(2), 259-281. https://doi.org/10.1080/13869795.2015.1032331

McGeer, V. (2014). P. F. Strawson's Consequentialism. In D. Shoemaker and N. Tognazzini (Eds.), *Oxford Studies in Agency and Responsibility, Vol. 2,* 64-92. Oxford: Oxford University Press.

McGeer, V., & Pettit, P. (2015). The Hard Problem of Responsibility. In D. Shoemaker (Ed.), *Oxford Studies in Agency and Responsibility, Volume 3,* 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,* 151-183. Oxford: Oxford University Press.

Millgram, E. (2014). Private Persons and Minimal Persons. *Journal of Social Philosophy 45*(3), 323-347. https://doi.org/10.1111/josp.12071

Mokhtar, T. H. (2019). Designing Social Robots at Scales Beyond the Humanoid. In O. Korn (Ed.), *Social Robots: Technological, Societal and Ethical Aspects of Human-Robot Interaction.* Switzerland: Springer Nature Switzerland.

Moro, C., Lin, S., Nejat, G., & Mihailidis, A. (2019). Social Robots and Seniors: A Comparative Study on the Influence of Dynamic Social Features on Human-Robot Interaction. *International Journal of Social Robotics, 11,* 5-24. https://doi.org/10.1007/s12369-018-0488-1

Nagel, T. (1986). *The View From Nowhere.* NY: Oxford University Press.

Nelkin, D. (2015). Psychopaths, Incorrigible Racists, and the Faces of Responsibility. *Ethics 125*(2), 357-390. https://doi.org/10.1086/678372

Nietzsche, F. (1874/2006). The Utility and Liability of History for Life. In K.A. Pearson and D. Large (Eds.) *The Nietzsche Reader.* Blackwell Publishing.

O'Neill, E. (2017). Kinds of Norms. *Philosophy Compass, 12*(5). https://doi.org/10.1111/phc3.12416

Pickard, H. (2020) Addiction and the self. *Noûs,* 1-25. https://doi.org/10.1111/nous.12328

Pickard, H. (2012). The Purpose in Chronic Addiction. *American Journal of Bioethics Neuroscience*, *3*(2), 40-49. https://doi.org/10.1080/21507740.2012.663058

Rabbitt, S. M., Kazdin, A. E., & Scassellati, B. (2015). Integrating socially assistive robotics into mental healthcare interventions: applications and recommendations for expanded use. *Clinical Psychology Review*, *35,* 35-46. https://doi.org/10.1016/j.cpr.2014.07.001

Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The Sources of Normativity: Young Children's Awareness of the Normative Structure of Games. *Developmental Psychology, 44*(3), 875-881. https://doi.org/10.1037/0012-1649.44.3.875

Railton, P. (2017). Moral Learning: Conceptual foundations and normative relevance. *Cognition, 167,* 172-190. https://doi.org/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. https://doi.org/10.1086/675876

Robinson, T., & Berridge, K. (2000). The psychology and neurobiology of addiction: an incentive-sensitization view. *Addiction, 95*(2): 91-117. https://doi.org/10.1080/09652140050111681

Ross, D. (2020). Addiction is socially engineered exploitation of natural biological vulnerability. *Behavioural Brain Research, 386*. https://doi.org/10.1016/j.bbr.2020.112598

Ross, D. (2017). Addicts and elephants: two varieties of diminished persons. (Unpublished manuscript) URL = < https://www.academia.edu/31618031/Addicts_and_elephants_two_varieties_of_diminished_persons >

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

Ryle, G. (1949/2002). *The Concept of Mind.* Chicago: University of Chicago Press.

Scanlon, T. M. (1998). *What We Owe to Each Other.* Cambridge: Harvard University Press.

Schlick, M. (1966). When is a Man Responsible? In B. Berofsky (Ed.), *Free Will and Determinism,* 54-62. NY: Harper & Row.

Schmidt, M. F.H., Rakoczy, H., & Tomasello, M. (2011). Young Children Attribute Normativity to Novel Actions without Pedagogy or Normative Language: Young Children Attribute Normativity, *Developmental Science*, *14*(3), 530-539. https://doi.org/10.1111/j.1467-7687.2010.01000.x

Shoemaker, D. (2015). *Responsibility from the Margins.* Oxford: Oxford University Press.

Shoemaker, D. (2011). Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. *Ethics, 121*(3), 602–632. https://doi.org/10.1086/659003

Smart, J. J. C. (1961).Free-Will, Praise, and Blame. *Mind, 70*(279), 291-306. https://doi.org/10.1093/mind/LXX.279.291

Smetana, J. G. (1993). Understanding of Social Rules. In M. Bennett (Ed.), *The Development of Social Cognition: The Child As Psychologist,* 111-141. NY: Guilford Press.

Smith, A. (2015). Responsibility as Answerability. *Inquiry 58*(2), 99-126. https://doi.org/10.1080/0020174X.2015.986851

Smith, A. (2013). Moral Blame and Moral Protest. In D. J. Coates & N. A. Tognazzini (Eds.), *Blame: Its Nature and Norms,* 27-48. Oxford: Oxford University Press.

Smith, A. (2012). Attributability, Answerability, and Accountability: In Defense of a Unified Account. *Ethics, 122*(3), 575-589. http://doi.org/10.1086/664752

Smith, A. (2008). Control, Responsibility and Moral Assessment. *Philosophical Studies, 138,* 367-392. https://doi.org/10.1007/s11098-006-9048-x

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*, 17–38. Oxford: Clarendon Press.

Sripada, C. 2017. Frankfurt's Unwilling and Willing Addicts. *Mind, 126*(503), 781-815. https://doi.org/10.1093/mind/fzw013

Sripada, C. 2016. Self-expression: a deep self theory of moral responsibility. *Philosophical Studies 173*(5), 1203-1232. https://doi.org/10.1007/s11098-015-0527-9

Sripada, C. 2015. Moral Responsibility, Reasons, and the Self. In D. Shoemaker (Ed.), *Oxford Studies in Agency and Responsibility, Vol. 3*, 242-264. Oxford: Oxford University Press.

Sripada, C. 2014. How is Willpower Possible? The Puzzle of Synchronic Self-Control and the Divided Mind. *Noûs, 48*(1), 41–74. https://doi.org/10.1111/j.1468-0068.2012.00870.x

Sripada, C. S., & Stich, S. (2007). A Framework for the Psychology of Norms. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind, Volume 2: Culture and Cognition*. New York: Oxford University Press, 280–301.

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>

Strawson, P. F. (1962/2008). Freedom and Resentment. In *Freedom and Resentment and Other Essays*. London, UK: Methuen, 1-25.

Su, N. M., Lazar, A., Bardzell, J., & Bardzell, S. (2019). Of Dolls and Men: Anticipating Sexual Intimacy with Robots. *ACM Transactions on Computer-Human Interaction, 26*(3), 1-35. https://doi.org/10.1145/3301422

Tomasello, M. (2009). *Why We Cooperate*. Cambridge, MA: MIT Press.

Tooby, J. & Cosmides, L. (1992). The Psychological Foundations of Culture. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. NY: Oxford University Press.

Turiel, E. (1983). *The Development of Social Knowledge*, Cambridge: Cambridge University Press.

van Wynsberghe, A. (2013). Designing Robots for Care: Care Centered Value-Sensitive Design. *Science and Engineering Ethics, 19,* 407-433. https://doi.org/10.1007/s11948-011-9343-6

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

Vincent, J., Taipale, S., Sapio, B., Lugano, G., & Fortunati, L. (Eds.) (2015), *Social Robots from a Human Perspective.* Swizerland: Springer International Publishing AG.

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,* 11-36. Oxford: Oxford University Press.

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

Watson, G. (2012). The Trouble with Psychopaths. In R. J. Wallace, R. Kumar, & S. Freeman (Eds.), *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon*, 307–31. Oxford: Oxford University Press.

Watson, G. (2004). *Agency and Answerability*. Oxford: Oxford University Press.

Watson, G. (1996/2004). Two Faces of Responsibility. In G. Watson (Ed.), *Agency and Answerability*, 260–86. Oxford: Oxford University Press.

Watson, G. (1987). Responsibility and the Limits of Evil. In F. D.Schoeman (Ed.), *Responsibility, Character, and the Emotions.* NY: Cambridge University Press.

Watson, G. (1975). Free agency. *The Journal of Philosophy, 72*(8), 205-220. https://doi.org/10.2307/2024703

Wilks, Y. (2005). Artificial Companions. *Interdisciplinary Science Reviews, 30*(2), 145-152. https://doi.org/10.1179/030801805X25945

Young, J. E., Hawkins, R., Sharlin, E., & Igarashi, T. (2008). Toward Acceptable Domestic Robots: Applying Insights from Social Psychology. *International Journal of Social Robotics, 1*. https://doi.org/10.1007/s12369-008-0006-y

Zawidzki, T. (2013). *Mindshaping.* MA: MIT Press.