Forthcoming in In T. Cyr, A. Law, & N. A. Tognazzini (Eds.), *Freedom, Responsibility, and Value: Essays in Honor of John Martin Fischer*. New York: Routledge.

Responsibility and Reasons-Responsiveness

Dana Kay Nelkin and Manuel Vargas
University of California San Diego

For any given topic in philosophy, there tends to be a multitude of theoretical options but only a handful of dominant views. For theories of free will and moral responsibility, reasons-responsiveness approaches are among that small set of the principal theoretical options. The core of these accounts, more or less, is the view that an agent is morally responsible when that agent, or some suitable element of that agent, can recognize and appropriately respond to reasons. In what follows, we explore some of the trade-offs for different ways of developing this basic picture, and we consider the resources of reasons-responsiveness views for addressing a variety of recent challenges that have been put to them.

Our focus is on two different approaches within the wider family of reasons-responsiveness theories, and the comparative appeal of each. In framing this as a consideration of two options within the reasons-responsiveness approach we do not mean to imply that these are the only options, or that the particular bundles of commitments that we identify cannot be mixed and matched in other ways. It is simply that these different approaches within the reasons-responsiveness family represent two of the better developed options; they are therefore particularly useful models for exploring the virtues of different ways of developing the general approach. It will surprise no one that we favor one package over the other. Still, the aim here is less about winning minds and hearts and more about casting some light on the diverse resources that this family of approaches has for informing our understanding of free will and moral responsibility.

We begin by outlining the two approaches to reasons-responsiveness, including the landmark version developed by John Martin Fischer, and a more recent set of convergent developments that we dub the Triton theory. With some characterization of those differences in hand, we focus on some of the choice points that separate these theories, focusing on three specific issues: (1) the role of circumstances, opportunities, or what is sometimes called the "ecology" of agency; (2) whether reasons-responsiveness is best understood as a feature of the agent or a mechanism of the agent; and (3) the role and significance of reflection.

1. Two pictures of reasons-responsiveness

The most important, influential, and richly developed account of reasons-responsiveness is undoubtedly John Martin Fischer's. Over the past 40 years or so, he and his sometimes co-authors have done more than anyone to show the power of the approach and its resources for addressing the various criticisms that have been directed against it. Although he was responsible for an extraordinary number of important ideas in theorizing about free will and moral responsibility, here we call attention to three ideas in Fischer's version of reasons-responsiveness that will figure in what follows.

First, although some have emphasized the possibility of reasons-responsiveness as a theory of free will, Fischer's primary deployment of the approach is as a theory of the ground or basis of moral responsibility, one whose tenability does not hang on fights over whether determinism rules out some or another notion of the ability to do otherwise (this is Fischer's idea of semicompatibilism). At the same time, reasons-responsiveness is not an account of culpability, blameworthiness, or accountability, so much as a story about a condition on these things—something he characterizes as responsibility, but which we will here characterize as responsible agency.

Second, Fischer showed just how capacious a "pure" reasons-responsiveness view could be. Some—Wolf (1987), for example, and more recently McKenna & van Schoelandt (2015)—have thought of reasons-responsiveness as a supplement or further condition on a mesh theory of the sort associated with Frankfurt (1971), Watson (1975), Bratman (2000).[1] However, one of Fischer's important insights was that the more parsimonious (i.e., non-mixed) version of the view was arguably every bit as appealing as a mixed view, without needing any extra machinery. At the same time, jettisoning mesh features for the theory of responsibility doesn't rule out other roles for identification or value-expression accounts. As Fischer has emphasized, even if a pure reasons-responsiveness is the right account of the basis of responsibility one might still hold that identification (or what have you) plays an important role in accounting for other important forms of agency (e.g., autonomous agency).

Third, where other theorists had been content to gesture at the idea of reasons-responsiveness, or invoked some general ability for rational self-governance, Fischer sought to provide a more detailed picture of what it was to have a rational ability. This was an especially important challenge to meet if the approach was to have serious credentials in a debate shaped by partisans of disputes about the metaphysics of free will. Only a couple of decades earlier compatibilist accounts had foundered on an inability to articulate the sense of 'can' that was putatively at stake in discussions of free will and moral responsibility. Absent some story about the sense of ability at stake in appeals to an ability for rational self-governance, the reasons-responsiveness project appeared to rely on stipulation to provide what was most needful. Fischer's strategy was to appeal to a sub-agential mechanism that has a Goldilocks modal profile—neither too lax nor too strict—which he characterized as *moderately reasons-responsive*.

[1] There have been diverse labels for this family of views. Among them: "identificationist," "self-expression," "valuational," "quality of will," "deep self," and "real self" views.

On Fischer's approach, reasons-responsiveness consists in two main components: the recognition of normative reasons (he dubs this *receptivity*) and the appropriate response to those reasons in intentions to act (or *reactivity*). To test for moderate reasons-responsiveness, we isolate a specific sub-agential mechanism (perhaps a particular bit of practical reasoning, although Fischer has always been careful to never commit himself to a specific model of the mechanism). Then, we ask whether it displays an "understandable" and "appropriate" pattern of regularly recognizing sufficient reasons, such that the agent recognizes how (interrelated) reasons "fit together" (Fischer and Ravizza 1998, p. 71). This is moderate receptivity.

If the agent has moderate receptivity, then we go on to test for the formation of intentions that are suitably responsive to sufficiently authoritative reasons for action (that is, reactivity, or what others have sometimes called *volitional control*). If the agent displays receptivity and any responsiveness to a sufficient reason to do otherwise, then according to Fischer and Ravizza, this suffices to show that the mechanism can react to *"any* incentive to do otherwise" (73). So, we have a moderately reasons-responsive agent when an agent's operative mechanism has a suitably robust, recognizably patterned, modal profile of recognizing sufficient reasons, and that mechanism also has any degree of sensitivity to sufficient reasons to do otherwise.

In addition to this picture of reasons-responsiveness, Fischer has argued for an independent requirement that the mechanism is the agent's own mechanism, where this is satisfied by the agent seeing herself in a particular way that amounts to taking responsibility for acting from that kind of mechanism (Fischer 2012, p. 187). The result is a picture of moral responsibility according to which it is "essentially historical," and not something that can be attained by reasons-responsiveness alone (Fischer and Ravizza 1998, p. 207).

Building on Fischer's trailblazing work, several philosophers have taken up the reasons-responsiveness framework and developed it in new and distinctive ways. In what follows, we focus on a family of views that share a distinctive set of commitments, including: (1) an agent-based (as opposed to Fischer's mechanism-based) approach to reasons-responsiveness; (2) what we might think of as an "ecological" picture of responsible agency according to which culpability for behavior is a function not just of agent-based properties but also of the context of action; (3) explicit scalarity in the idea of degrees of responsiveness and corresponding degrees of excuse; and (4) accommodation of the idea that exercises of responsible agency provide some basis for deserving blame. Given that these commitments are especially prominent in the individual and collective work of David Brink, Dana Nelkin, and Manuel Vargas, we'll characterize this set of commitments as the Triton theory of reasons-responsiveness.

In characterizing the Triton theory, we do not mean to imply that its proponents agree on the most perspicuous way to regiment the central commitments. Moreover, the individual theories differ in some important ways. For example, one account is more elegantly stated, another employs particularly clever arguments, and the third is strangely compelling. Despite these and less serious differences, we

believe that the overlap of commitments is sufficiently robust that it makes sense to speak of a shared theoretical framework.

For the sake of clarifying some points of convergence among Triton theories, we offer a brief translation scheme for the various extant regimentations of the convergent ideas. Then, in the remainder of this chapter we consider several issues on which the verdicts of the Fischer-style and Triton theories come apart, and we consider the implications.

The core of the Triton theory holds that for an action to be morally blameworthy in the accountability sense, and in a way not derivative on some prior episode exercise of agency, the agent must act wrongly, be an agent of the relevant sort, and be in circumstances the nature of which makes it reasonable to demand compliance with norms enjoining one to not act wrongly. Putting aside the condition of acting wrongly, the central features are an account of the required form of agency and the nature of circumstances. Tritons are rational capacitarians, which is to say reasons-responsiveness theorists about the kind of agency required for (non-derivative) responsibility.

As in Fischer's account, Tritons recognize two main components in the agent-based part of the account. First, agents must be able to recognize relevant reasons or considerations. (The terminology or governing category employed here has varied among Tritons, including "recognitional capacity" (Vargas 2013); "cognitive capacities" (Brink and Nelkin 2013); "cognitive competence" (Brink 2021); and "cognitive control" (Vargas forthcoming).) Second, agents must be able to suitably respond to those considerations. Triton theorists have variously characterized this as a "volitional capacity" (Brink and Nelkin 2013; Vargas 2013); "volitional competence" (Brink 2021); and "volitional control" (Vargas forthcoming). Finally, the conjunction of those powers has been characterized in different ways—for Brink and Nelkin (2013), these abilities jointly constitute normative competence, for Vargas (2013), they constitute free will, understood in a revisionary way.

The agential features thus far identified are all analogs of the core picture set out by Fischer. However, Triton theories differ from Fischer-style accounts in their ambitions for they seek to offer an account of blameworthiness, and not just an account of responsible agency.[2] To do that, they invoke

[2] For Fischer, reasons-responsiveness is an account of responsible agency, but not of culpability (or what is sometimes called accountability blameworthiness). Fischer and Tognazzini (2011) specify that the gap between reasons-responsiveness and culpability is closed when one establishes that some wrongful action lacks justification or an excuse. They regard justification and excuse as occupying what they call "The Space Between" attributability and accountability although they only gesture at how an account of excuses might go. Their example concerns a case of duress, which they construe in terms of the difficulty of doing the right thing. Coates and Swenson (2013) have tried to extend Fischer's account of reasons-responsiveness in a natural way to capture the idea of difficulty, appealing to the scalar aspect of reasons-responsiveness so that the more different circumstances would have to be for one to act on the reasons there are, the more difficult and the less blameworthy. To our knowledge, Fischer has not endorsed this approach, and we believe that there is good reason not to. Without recognizing a distinctive role for actual situational factors, reasons-responsiveness of mechanisms cannot correctly capture cases of difficulty (Nelkin 2016). This means that the framework needs a different kind of supplementation to capture the excuse of difficulty. Moreover, the difference between cases of difficulty that excuse and cases of akrasia is not obviously accounted for in terms of difficulty alone. One might worry that in the absence of an account of what fills The Space Between, it is unclear how

the idea of a further, ecological feature beyond the agent and the wrongfulness of the action. That is, they appeal to the nature or quality of the situation or context in which the agent operates. Notably, the ecological component is not simply a further and entirely independent condition. Rather, its significance is relational and interactive, in that changes to the ecology can alter the normative significance of an agent's intrinsic features.

Tritons have had different ways of characterizing that relationality. On one regimentation, ecological features bear on whether an agent has responsibility-relevant control (as in the account of circumstantialism in Vargas 2013). On another regimentation, control or the abilities that matter for responsibility are construed as a purely agent-restricted phenomenon, but the opportunities afforded an agent by context alters the moral significance of those abilities (as in Brink and Nelkin 2013; Brink 2021, p. 91). These latter regimentations are "fair opportunity" or "quality of opportunity" views. However, the shared idea across all these accounts is that situations can vary in their "degree of congeniality" for various normative demands, and this matters for culpability (Nelkin 2020, p. 208).

We think these different framings of a core set of ideas are mostly bookkeeping differences rather than substantive differences. In canvassing the ecological aspect of the Triton theory, we do not mean to imply that an ecological approach is unique to it. One might accept an ecological account without accepting agent-based reasons-responsiveness, or without going in for scalarity about abilities and culpability, or without accepting the centrality of desert. And, of course, there are differences between individual Triton theories. For example, the principal proponents of the Triton theory differ in how they cash out the modal properties invoked by their respective accounts, in the broader normative commitments presumed in accounting for responsibility, in some methodological presumptions, in their commitments to various forms of realism, and so on. In the present context, though, these differences are mostly immaterial, so we won't belabor them.

In the sections that follow we consider the resources within (or readily available to) these two approaches to reasons-responsiveness as they bear on a trio of interlocking choice points that face reasons-responsiveness views. The first concerns the grounds for adopting the ecological component that is especially prominent in Triton theories. The second concerns reasons for favoring an agent-based or mechanism-based account of reasons-responsiveness, and the third takes up the role of reflection for responsible agency. Throughout, we are less concerned to settle particular philosophical debates than we are calling attention to some differences in resources, motivations, and capacities available to different ways of developing the idea of reasons-responsiveness.

much work the reasons-responsiveness element of Fischer's account does in explaining the kind of culpability that tends to motivate an interest in free will and moral responsibility. We return to some of these issues in §2 and §5.

2. Ecology

Above, we noted that a distinctive feature of the Triton theories is their explicit commitment to what we might think of as an ecological picture of responsible agency. That is, beyond questions about the agent's rational capacities, the Triton theories appeal to facts about the ecology of action—that is, the context, circumstances, situation, or opportunities, which may include things as diverse as social norms and practices, institutions and their affordances, the arrangement of material resources, and so on. Here, we outline some general motivations for taking on these commitments.

One reason for thinking it is important to explicitly say something about the role of circumstances is that it seems to matter relatively directly for any capacitarian theory. In natural language, talk of capacities tends to leave tacit the boundedness of (or delimitations on) the capacity. The ability to speak a language tends to presume contexts where one is conscious; the ability to swim presumes that one is not in a vat of acid, and so on. One motivation for ecological accounts is that they make explicit what is otherwise at risk of being overlooked: that the impediments or demands of contexts alter what it makes sense to demand of agents, and thus, what faults or credit we attribute to agents. Once you acknowledge the boundedness of capacities, though, you need some account of the bounds of those capacities, or at least, an account of important defeaters for the presumption that the capacity is present in everyday actions.

Candidate cases of culpability under conditions of oppression, domination, or deprivation are especially salient instances of where ecological theories can give responsibility theorists important resources. On non-ecological accounts, it can appear that the only way such conditions affect culpability is via impairments to cognitive and volitional control. Ecological theorists, though, can appeal to the idea that differences in context can alter the reasonableness of demands to exercise one's agency in particular ways. Alternatively, the ecological theorist can insist that these conditions alter whether it is appropriate to attribute a responsibility-relevant capacity to a person.[3] Accounts of responsibility that do not appeal to these extrinsic features of agents have no basis for capturing the thought that the external conditions of agency have a direct bearing on culpability. Intuitively, some of those external conditions are such that make it more difficult for agents to act well for good reasons.

Another range of cases that call out for a picture of environmentally conditioned control concerns coercion or duress. (For present purposes, we'll treat these as interchangeable.) One way these cases might work is by altering the psychological functioning of agents. If someone threatens you with "your money or your life!" this might induce panic or cause some other effect that swamps the ordinary capacities of an agent to recognize and respond to moral considerations. Yet, the significance of duress doesn't seem to require that this is what is happening. It might be that you are fully capable of

[3] We hasten to add that the effect need not always be one of mitigation; unjust conditions may heighten some people's sensitivity to injustice and/or may invigorate their inclination to address it (Vargas 2018a).

recognizing and responding to sufficient reasons to act otherwise, but the incentives in the actual sequence don't give you reason to do so. Still, these external features seem to matter.

One appealing feature of ecological approaches to responsibility is their continuity with broadly ecological approaches to agency more generally. Those approaches have included discussions of the ecology of practical reason (Morton 2011; 2022), autonomy (Mackenzie and Stoljar 2000; Mackenzie 2015), and the social meaning of acts (Bierria 2014).[4] One can also find continuities with the criminal law's distinction between capacity and opportunity, and the philosopher's thought that these things interact in a way that matters morally (Brink 2021).

Again, none of this is to say that non-Triton theories are precluded from making use of these ideas, or that they lack independent theoretical resources for capturing some of these phenomena. Still, the explicit invocation of these considerations and their deployment in theoretical contexts is a distinctive characteristic of Triton theories, and a natural question to press to Fischer-style views is whether such accounts are inclined to take on board these commitments, or instead, whether they eschew them. If the latter, we might wonder how the phenomena canvassed here might be addressed in a framework that only grounds culpability in cognitive and volitional control.

3. Agent vs. mechanism

Is reasons-responsiveness best understood as primarily a property of a sub-agential mechanism and only derivatively of an agent, or instead, as a property of the agent? Fischer-style accounts claim the first, Triton theories the second.

The difference is not merely notational. For agent-based approaches, there may be no single individual mechanism that realizes reasons-responsiveness. Instead, it may be a property of an entire system of interacting dispositions, information processing, and contingent attunements to an environment or relations in it. On this approach, reasons-responsiveness is more like intelligence or charisma, in that it is an agent-level property that is often a product of a diverse variety of interacting, lower-level phenomena.

In contrast, the mechanism-based approach identifies a specific sub-agential process or functioning—a mechanism, in Fischer's terminology—that is the bearer of reasons-responsiveness. On this model, reasons-responsiveness is more like the property of genetic mutation, that is, fully explicable at the lower-level (that is, sub-agential) phenomenon, even when it has outsize effects for the organism. One piece of evidence that mechanism talk is intended to mark this difference becomes apparent when we consider why there was any benefit to mechanism talk in the first place. If the organizational level at which reasons-responsiveness attaches is immaterial, there would have been no need to invoke mechanism talk at all, or to focus on whether there are different mechanisms operative across cases.

[4] These issues have also had an especially robust significance in the history of Latin American philosophy (e.g., Vargas 2020; 2022) and in the work of Latina feminist philosophy in the U.S. (e.g., Schutte 1998, Lugones 2003).

We're inclined to think that there are some modest reasons to presumptively favor an agent-based picture.[5] First, consider the pedestrian fact of ordinary language and practice: we blame people and not mechanisms. Were one to respond to an accusation of culpability by averring to the modal properties of some agential sub-system—"Don't blame me! My deliberative mechanism was janky"—this would strike us as either an effort to dissociate from one's wrongdoing or an ungainly appeal to 1980s-style cognitive neuroscience reductionism.

One explanation for the apparent oddity of that thought is that the property at stake is a systemic one, something located in the functioning of the agent as a whole, and not just in the state of a specific part. The significance of this difference most readily emerges when one considers that in holding fixed a given mechanism and checking for its failure in other worlds ignores the effects of other, potentially redundant (in the actual sequence) subagential mechanisms. Those other mechanisms can alter the agent's overall modal profile in a way that comes apart from the modal profile of the mechanism we are holding fixed.

For example, suppose someone acts out of self-deception, and that the relevant mechanism is not a moderately reasons-responsive one. It still seems an open question whether the agent could have worked around it, perhaps by deploying some other, more reasons-responsive mechanism. This fits well with the thought that we do not automatically excuse people for acting on such mechanisms, as we ought to if acting on a reasons-responsive mechanism were necessary for responsibility (Nelkin 2012). Similarly, the performance of a given mechanism in the actual world might be masked or finked by other aspects of an agent's psychology, e.g., mood disorders like depression, grief, exhaustion, and so on (Nelkin 2016). Perhaps a more detailed account of mechanisms and their individuation would help alleviate these concerns, although it is not intuitively obvious to us how any account that isolates a single sub-agential system will be able to entirely avoid this sort of concern.

A second consideration for favoring an agent-based account is that, plausibly, the reasons that matter for culpability are reasons *for the agent*, not the reasons for or from the standpoint of a subagential mechanism. For example, people with alien hand syndrome or various unusual pain syndromes show sub-personal responses to reasons or candidate reasons that aren't the agent's reasons. So, too, with the

[5] We note that the main motivation for Fischer's mechanism-based account is that he takes it that it can better accommodate the intuitions generated by "Frankfurt cases." These are cases in which one intuitively acts freely and responsibly even though there is another agent "waiting in the wings" who would (counterfactually) intervene if one started to act in a different way that one actually does. Cases of these sorts led Frankfurt to conclude that the ability to do otherwise is not necessary for acting freely and responsibly. Fischer agrees that the agent is not reasons-responsiveness in such cases, but he holds the mechanism on which she acts is (Fischer and Ravizza (1988), p. 38). While this is an ingenious move in response to Frankfurt cases, we do not believe that things are so dire for the agent in such a case as Fischer thinks. Although we do not develop the point here, we believe that building in the ecological component discussed earlier provides the tools to account for an agent's capacity in the relevant ways. For arguments that agents can do otherwise in some relevant sense, e.g., Wolf (1990), Vihvelin (2004), and Nelkin (2011). For an illuminating and systematic discussion of Frankfurt cases and reasons-responsiveness, see McKenna (2022).

apparent reasons of hunger and pain. To cabin off some part of the agent as the mechanism is to abstract away from the systemic significance of the reasons at stake in our moral evaluation of agents as candidates for culpability. (One might add that this is why it seems optimistic to think science is going to help us solve puzzles about when high level failures of reasons-responsiveness are culpable.)

Last, there are a cluster of familiar challenges to an approach that invokes mechanisms as the bearer of the reasons-responsiveness property, including questions about how to identify the relevant mechanism and how to individuate mechanisms. These are challenges that are unique to a mechanism-based theory, over and above the more general challenge of whether reasons-responsiveness theories that invoke talk of capacities and abilities can provide a satisfying metaphysics of those modal notions.

In the remainder of this section, we call attention to a recent challenge that Chandra Sripada (2019) has put to reasons-responsiveness theories. We think the challenge can be met by Triton Theories. It is less clear whether there is a comparably effective set of replies available to Fischer-style views, precisely because of their reliance on a mechanism-based story. However, we raise Sripada's challenge not because we endorse it as decisive against Fischer-style views, but rather by way of invitation for proponents of mechanism-based accounts to say how they think a response should go, given their theoretical commitments.

Sripada's Fallibility Challenge works by way of an example that purports to show that, given a reasons-responsive theory of responsibility, an agent must be held culpable for something that is intuitively not a case of fault. (Sripada calls this the "Fallibility Paradox" but one might contest whether it is a paradox; we say more about this below.) Our modest reformulation of the example, drawing from Sripada (2019, pp. 235-6), is as follows:

> FEI: Fei performs 1000 trials of a cognitively demanding and time-pressured task. She is incentivized to succeed, as her rate of success affects the amount donated to a worthy charity about which she cares deeply. Consequently, she exerts tremendous effort to be as accurate as possible. Impressively, her performance is at the upper limit of human success: she makes only four errors in a thousand trials, giving her a success rate of more than 99.5%.

Consider any arbitrary trial of the task. Fei acts intentionally in selecting answers, and she aims at getting the correct answer. Sripada assumes that (1) in so acting Fei plausibly operates from a reasons-responsive mechanism; and (2) the mechanism from which Fei operates is the same across trials. If that's right, and assuming there are no defeaters, then it looks like any reasons-responsiveness theory should hold that for any given trial Fei is responsible (to be credited) when she succeeds, and that she is also responsible (blameworthy) in the four instances when she fails.[6]

[6] One might dodge this conclusion if one insisted that one must be perfectly responsive to reasons to count as reasons-responsive enough for responsibility, but this reply has a cost of its own. It entails that any instance of failure is its own exculpation, for failure would provide authoritative grounds that the agent is not reasons-responsive.

The problem for reasons-responsiveness theorists is that this result seems counterintuitive. After all, Fei is effortfully trying to do the right thing and her performance is at the upper limit of human abilities. Indeed, as Sripada goes on to argue, there are some in-principle reasons to think that people cannot ordinarily avoid making some errors in a task with the parameters he describes, even when exerting maximal effort. So, reasons-responsiveness theories face a problem.

We have three general observations about the scope of the Fallibility Challenge. First, if it is a problem, it is primarily a problem for "pure" reasons-responsiveness views, and not views that mix reasons-responsiveness with other conditions.

Mixed views take reasons-responsiveness as a condition on culpability but hold that there are also other conditions. There are at least two theoretical roles for other conditions. One is in the specification of what makes someone a responsible agent—that is, an apt candidate for responsibility practices. The other is in the conditions for blameworthiness. McKenna and Van Schoelandt's account of a "Hybrid-Mesh-RR" view of free agency suggests the former, as does Wolf's (1987) earlier proposal to see reasons-responsiveness (what she calls "sanity") as a supplement to an identificationist or Deep Self view. The proposal by Vargas (2013, p. 180-181; 237) is explicitly an instance of the latter, with a reasons-responsiveness account of responsible agency and a quality of will story about the general character of culpability norms (which are themselves partly given by features of the interaction of agency and ecology). And, as we will see below, the resources in this latter approach plausibly generalize to any ecological view.

The reason the Fallibility Challenge is primarily a problem for pure views is that mixed views can appeal to the non-reasons-responsive element of the theory to account for why Fei was not culpable in instances of failure. So, depending on the particulars of a given mixed view, Fei's non-culpability for failure cases can be explained by her failure to act with poor quality of will or her not identifying with the failure. In short, Sripada's claim that his challenge "applies to any version of a reasons-responsiveness view" (2019, p. 245) cannot be sustained.

Still, there is at least a genuine prima facie challenge for "pure" reasons-responsiveness approaches, and for any mixed view that cannot explain Fei's non-culpability by appeal to whatever additional conditions they contain. Fischer's account is a mixed one insofar as it includes the Taking Responsibility condition; that is, in addition to acting on a reasons-responsive mechanism, to be responsible, the agent must have taken responsibility for it by viewing themselves as an apt target of blame and praise for actions or omissions that issue from it. So, there is a second condition that is a potential resource for addressing the challenge. But it appears that it cannot help here. Presumably Fei has taken responsibility for the mechanism or else she wouldn't be responsible in the successful cases either. So that means she would be responsible in all the cases.

Sripada initially frames his example in terms of a mechanism-based theory, but he holds that it applies equally well to agent-based accounts. In the remainder of this section, we consider whether it

does. Predictably, we think it doesn't. If that's right, it gives us a further reason to favor agent-based accounts over mechanism-based accounts.

Sripada claims that the basic problem contained in the Fallibility Challenge obtains with any view of reasons-responsiveness that permits capacity failures, but we believe that this is because he is not considering the right kind of agent-based account. Sripada understands the difference between mechanism-based accounts and agent-based accounts as follows. Mechanism-based views assess the responsivity to reasons of the *operative subset* of the complete set, *Y*, of processes that make up an agent's psychology. In contrast, "[a]gent-based views differ from mechanism-based views in that they say we need to assess the responsivity to reasons of Y itself (all the psychological processes of the agent) rather than a subset of *Y*" (p. 245). In this way, it is easy to see how Sripada would conclude that agent-based accounts immediately inherit the challenge facing mechanism-based ones. After all, on this view, agents are just (larger) aggregates of mechanisms than those that are at the heart of mechanism-based views.

We are unpersuaded that the aggregative picture is the best way to understand agent-based accounts. Even if the agent were constituted by the various mechanisms, the agent is not reducible to the aggregate of them, nor is the agent's degree of reasons-responsiveness just the aggregated or average responsiveness of the mechanisms. This is, in part, because structural relations among the mechanisms, and the possibility of their interaction, can affect the degree of reasons-responsiveness of the agent as a whole. (We saw a version of this point above, regarding the significance of overlapping but distinct mechanisms that jointly give the agent a different modal profile than the profile of any individual mechanism.) We believe that once we have this understanding of the agent in hand, together with other elements of the Triton approach, such as the ecological one that interacts with the relevant agential features in any given situation, there is room for a different kind of response to the Fallibility Challenge.

One place the agent-based theorist is going to focus on are the failure cases, where Fei gets the wrong answer. First, let's suppose that Fei is trying as hard as she can to get it right in the failure cases. Whether because of distraction, the degradation of attentional resources, or some combination of these things, if there is nothing else the agent can do, no arrangement of her attention or volitional control that makes a difference, this begins to look like the absence of an agent-level capacity in that instance, rather than a failure of capacity.

Sripada suggests that failure gets going because of unavoidable "noise-based errors" that are due to the stochastic nature of the underlying neurological process (p. 242). This is what explains why some cases are error cases and others are not. If that's right though, in precisely those instances where Fei is trying as hard as she possibly can and her goals are frustrated by a noise-based error, it looks like she has an (agent-based!) explanation for why she isn't culpable: there was nothing more she could do, *qua* agent.

The metaphysics of capacities is a messy business, but the broad avenues for response are clear enough. Depending on the details of one's theory, Fei will have either an excuse or an exemption. She

has an excuse if one's theory holds that she has a capacity to get the right answer, but it is unreasonable to demand that she get the right answer. She has an exemption if one's theory holds that in those cases she lacks the capacity to get the right answer.[7] In either case, though, Fei is not culpable for her failure. By stipulation, she's trying as hard as she can and that's still not enough to prevent the performance failure. So, it is unreasonable to demand that she try any harder, and consequently, there is no basis to hold her culpable for failures.

Now let's suppose that Fei is not trying as hard as she can to get it right in the failure cases, and that her trying harder would make a difference to her performance in those instances of failure. If she's culpable for those failures, that's the explanation for why. But is she culpable for those failures? We could decide that she is. After all, she has a capacity to try harder, there is some reason to do so, and she doesn't. We think some agent-based theorists will be satisfied with this analysis. Still, there is potentially more that can be said by an ecologically-inclined agent-based theorist. In particular, at least some ecological approaches will have reason to think we might distinguish between different kinds of performance failures, even holding fixed Fei's physiological facts.

Recall that what the agent-based theorist wants to know is something about the rational status of the agent, at least with respect to the reasons that bear on the issue under consideration. On ecological agent-based theories, we also need to know something further still: the relevant context/circumstances/opportunities. In both instances—in focusing on the agent and on the ecology—the reason for doing so is that such things bear on what is reasonable to demand of those agents. To address such questions, it is not enough to provide rich accounts of mechanisms and metaphysics. A story about moral responsibility needs a story about the norms that properly bear on those things. That's why agents and ecologies matter: they are among the things that ground explanations of why some capacity failures are culpable and others are not.

Go back to Fei's four failures. In each of the instances when she didn't fully exert herself, we can still ask whether we can demand that she have exerted herself more fully. And here, context matters a great deal. If it were a one-off instance, and the stakes were sufficiently high, then we might think that we could demand greater effort. If she failed when giving it her all (and assuming there was no prior culpable failure on which to ground her culpability for failing in the task—for example, knowingly taking drugs that would impair her attention right before undertaking the task), then she is plausibly non-culpable. However, in a case where the stakes are relatively low, where performance is iterative, where finite cognitive resources are nearly exhausted, and when agent-level control is subject to some inevitable cognitive noise, it does not seem reasonable to demand successful performance in every trial.

[7] Some theories of responsible agency tend to portray it as a quasi-permanent status ordinarily achieved by adulthood such that exemptions (or absences of responsible agency) are rare. However, some theories hold that responsible agency is a relatively "patchy" property that comes and goes across contexts (Vargas 2013). Patchiness about responsible agency/exemptions is compatible with there also being excuses available to responsible agents.

The responsibility theorist—plausibly, a species of *moral* theorist—should want a theory that is sensitive to such considerations. That kind of account requires looking past the mechanism and to our normative interests in agents and their performance in the socially and normatively significant features of context—including the situational difficulty of the task.

That we can and do regard such features as significant is apparent from the subtlety with which we adjust our normative evaluations of agents under an array of circumstances. If the championship is on the line, and the team's best striker fails an open shot that nearly everyone in the sport—the striker included—thinks is culpable, then most of us will be inclined to agree. That's compatible, though, with thinking that it is non-obvious whether the shooter is at fault if it is the team's third string keeper who misses the same shot. At the same time, we recognize that the relevant standards shift if the context is iterative, where exhaustion and diminishing attention and skill are factors. And we recognize that depending on whether the target of evaluation is an individual performance or a set of performances, our evaluations can change yet again.

We concede that there can be an air of the paradoxical about all of this. That it would be unreasonable to demand that a striker succeed at every game-winning shot over a career is compatible with thinking that in any given instance, one is culpable for missing that particular shot. As in the lottery paradox, agglomeration can and perhaps should be resisted (Klyberg 1961).[8] In general, the normative evaluation of sets of actions need not invoke the same standards that are used to evaluate any individual performance of some action. (This is a point that, coincidentally, emerges in at least some instrumentalist accounts of responsibility, as in Vargas 2013.) Instead, our norms of assessment are sensitive to roles, interests, and expectations in a way that cannot be reconstructed with any amount of attention to sub-agential mechanisms.

The theme we have been pressing is that a satisfactory theory of culpability must be sensitive to the context in questions—whether an instance or set—and that the context matters for what we can reasonably expect of those agents. Indeed, altering details of the Fei case may make this especially clear. Let's suppose the failure cases were each cases where Fei failed to maximally exert herself. It seems to matter to our assessment of those failures if, for example, the following conditions also obtain: Fei volunteered for the task and the stakes were the ongoing existence of alpacas and Australian Shepherd puppies. In that case, we (and Fei) would have reason to regard her failures differently than if Fei were simply an unremarkable undergraduate who happened to volunteer for a psychology experiment with only a modest donation to a charity of dubious value.

---

[8] Sripada explicitly rejects an agglomerative reading of the Fei case, and his principal concern seems to be the identification of a case where a mechanism is reasons-responsive, but the action is intuitively not culpable. We've tried to follow that framing of the Challenge, but it does leave it a bit opaque why there is a paradox here, as Sripada's presentation emphasizes. Absent agglomeration or some similar principle, there is no obvious contradiction or apparent contradiction in any substantive sense. For this reason, it has seemed more apt to us to frame it as a challenge, rather than a paradox.

Elsewhere, Triton theorists have endeavored to say something about the normative basis of our responsibility norms, and the principled basis for the thresholds that emerge (e.g., Brink 2021; Nelkin 2016; Vargas 2013; 2018a; 2021; forthcoming). We won't try to recapitulate all those moves here. The more modest point is simply that the force of the Fallibility Challenge relies on reasons-responsiveness being construed in mechanism-based terms, and that agent-based and ecological accounts have additional resources to explain our varied intuitions about cases. Indeed, such accounts provide us with a kind of diagnosis about why, if one focuses only on mechanisms, it might seem hopeless to explain our varied intuitions across cases where the operative mechanism has the same modal profile.

The distinctive claim of ecological theories is that culpability is a function of the agent in circumstances, or the normative significance of the agent so acting given some opportunities. Consequently, we can't read off culpability from a performance failure of an isolated subagential system. By the lights of a Triton theorist, this would be akin to thinking you could decide whether the ball going over a fence counts as a homerun without knowing whether there is a game being played and where the field markers are.

This then gives us a diagnosis for why the Fallibility Challenge seems forceful for mechanism-based theories: it calls attention to the limited resources available for grounding the normative authority of blame in a mechanism-based approach. However, Triton theories have an additional resource for explaining the difference. On Triton theories, culpability isn't settled by just the state of a single sub-agential system, or even the state of the entirety of an agent, but the normative significance of the agent and the behavior in a set of circumstances.

4. Reflection

Reasons-responsiveness is a powerful capacity, but interestingly it is not one that is unique to humans, or even "persons". What seems to set persons apart is a capacity for reflection. In Harry Frankfurt's words, it is a second-order "capacity for reflective self-evaluation" (1971, p. 7), and in Christine Korsgaard's, it is a matter of having "reflective distance" that allows us to be

> "...active, self-directing, with respect to our beliefs and actions to a greater extent than the other animals are [insofar as] we can accept or reject the grounds of belief and action that perception and desire offer to us (2009, p. 32)."

This capacity to step back and assess our own beliefs and motives not only marks a distinction between "persons" and "non-persons" in an important sense, but also seems key to explaining why only reflective beings can be morally responsible agents.

This gives rise to a challenge for reasons-responsiveness views in all versions. If reasons-responsiveness does not itself entail reflective capacity, then it would seem it cannot provide a sufficient

condition for responsible agency. Any reasons-responsiveness account would need supplementation by a condition that captures reflective capacity, or it would need to explain why an additional condition is not necessary. Each approach comes with its own further challenges. On the first, we acquire the need to explain why such a general capacity is needed given that at least much of the time when we act in perfectly responsible ways that make us eligible for praise and blame, no such capacity is even available to us. On the second, we must explain away what has seemed a fixed point to many (including Frankfurt and Korsgaard).

To bring out the challenge for the first approach, we can survey the mountain of evidence that we are often unable to even recognize our own operative motives and beliefs, even when we appear to be apt targets of praise and blame. To take just one example, in an experiment in which researchers tested factors that affect whether people working in an office contribute to the collective coffee fund, they found that the mere presence of a drawing of two eyes next to the collection bowl increased contributions. But none of the subjects expressed awareness of this as a factor in their acting (Bateson et al (2006), discussed in Doris (2015). Dual process theory divides human cognition into two kinds of processes: system I, which is automatic, fast, effortless, and unreflective, and system II which is conscious, effortful, deliberative, and employing "executive control". (See Doris p. 50 for a summary.) Although this distinction is too simple in various ways (e.g., there is interaction between the systems), recent thinking about the two processes is that both can be intelligent and reasons-responsive. For example, we can even track moral reasons when in system I. Railton (2009), for example, describes cases in which people instantly (in this case, automatically, effortlessly, and without reflection) respond to the sight of another in need. In fact, it is sometimes claimed that in such cases, reflection can only interrupt the flow and undermine the skillful response.

In addition to this argument based on empirical evidence, there is an intriguing argument that it would be impossible to be reflective about *all* of our actions and omissions that we seem to be responsible for. For our reflective activity itself seems to be something we are responsible for, and yet if we were required to reflect on that very activity in order to be responsible for it, we would immediately encounter an infinite regress of a sort. As Arpaly and Schroeder (2012) have argued, one would have to reflect on our reflection about the reflection, and so on. Therefore, actual reflection cannot be required in order to be responsible for our actions. But then we need to know why the mere possession of the capacity should be required, especially when in many particular instances, the capacity does not even seem available to be exercised.

The challenge for reasons-responsiveness theorists is thus complex. On the one hand, reasons-responsiveness alone does not seem to entail a reflective capacity, which seems intuitively to be required for responsible agency. But simply taking such a capacity off the shelf to add to the account brings further explanatory burdens.

Fischer faces the challenge from empirical results directly when he acknowledges John Doris' argument of just this kind against Reflectivism, the view that "the exercise of human agency consists in judgment and behavior ordered by self-conscious reflection about what to think and do" (2015, p. x). Fischer writes that it "has intrigued and vexed me for a long time" (2018, p. 250).

Perhaps Fischer's account already contains something that entails reflective capacity of a particular sort? If so, then we might be able to address the challenge in a way that vindicates the appealing idea that reflective capacity is essential to responsible agency while also explaining how it can be that the capacity needn't be accessed (or even accessible) on all occasions on which people are responsible and even blameworthy. Fischer seems to suggest that the challenge for reflectivism, consisting in the mountain of evidence described earlier, is simply a challenge for reasons-responsiveness itself. And indeed, he concludes, "I really do think that moral responsibility requires reasons-responsiveness. So, I would conclude, under the envisaged circumstance in which I were convinced of the skeptical results [favoring the rejection of Reflectivism], that we are not morally responsible (p. 252)." But if we are correct in our interpretation of at least some of the experimental work, there is a special threat to *reflective* agency that leaves reasons-responsiveness untouched. In cases of acting in flow, for example, we can be acutely reasons-responsive without being reflective. Reflection is even thought to undermine our responsiveness in such cases. This means that reflection and reasons-responsiveness do not stand and fall together, and that we must inquire further about whether there is an essential role for reflective agency, and, if so, what it is.

But there is more to Fischer's rich account, and it might be thought that this is a place where it has an advantage over the Triton approach. In particular, Fischer and Ravizza's historical Taking Responsibility condition provides a natural candidate to locate a recognized role for reflection. Recall that to truly act with guidance control, on Fischer's view, one must act on a reasons-responsive mechanism that is *one's own*, where it is one's own only if one has taken responsibility for it by viewing oneself as a fair target of the reactive and blaming and praising attitudes when one acts on the relevant mechanism (1998, p. 210). This condition clearly presupposes a reflective capacity: one must be able to step back and take second-order attitudes toward oneself and one's agential mechanisms. This is a rich resource for meeting the challenge, and one unique to Fischer's version of a reasons-responsiveness view. Does it work? We do not think it ultimately succeeds, but seeing why is instructive.

Before going further, we set aside one immediate issue, namely, the prominent role of mechanisms in this part of the account. As Fischer (together with co-author, Ravizza) is aware, there are difficulties in individuating mechanisms in this context, and, as we have argued earlier, there is good reason to think of reasons-responsiveness as attaching to the agent and not the mechanism. Even if these problems could be side-stepped—perhaps instead of taking responsibility by making a mechanism one's own, one can take responsibility for one's actions in general—obstacles remain. Here, we highlight one

such obstacle, one that we believe reveals a new kind of objection to the Taking Responsibility condition based on our earlier discussion of the ecological/situational factor in responsible agency.

The challenge begins with the observation that there seem to be straightforward counterexamples to the Taking Responsibility condition on responsible agency. It is often the case that people sincerely disavow responsibility and yet we continue to think that they are responsible and even blameworthy for their actions. To their credit, Fischer and Ravizza (1998) consider an objection of this sort to their view. They respond by making vivid the supposition that a creature really has no sense of themselves as free and responsible. Such a creature would seem to be like Galen Strawson's "natural Epictetans," floating through life, never pausing to think of themselves as free and responsible or as having control. Intuitively, such creatures are so unlike us, and it is natural to reject the idea that they are responsible agents (pp. 221-223).

Perhaps this is the correct reaction to the Epictetan thought experiment. But the natural Epictetans are different from us in several ways, only one of which is their failure to take responsibility for a particular reasons-responsive mechanism. So, it is not the best test case of their account and the essential role of the Taking Responsibility condition. Consider instead people like us who *in general* have reflective capacities, but who, on a given occasion, reject the idea that they are responsible and blameworthy. To pick up on our ecological theme, consider people who act wrongly and from a moderately reasons-responsive mechanism, but the situation in which they find themselves is so challenging and uncongenial that they–rightly–think of themselves as completely excused. If they disavow responsibility, why should we think it important to determining whether they are responsible or not that they earlier took responsibility in a blanket way for any action that issues from the operative mechanism?

We see two possible responses that might be made on behalf of the idea that responsible agency really does require that the agent have taken responsibility. First, Fischer might at this point appeal to the aforementioned distinction between responsibility on the one hand, and blameworthiness or culpability on the other (Fischer and Tognazzini 2011). He could deny that seeing oneself as excused entails seeing oneself as non-responsible. This is a fair point, but we believe that its significance in this context has been underexplored. If reasons-responsiveness is *only* an account of responsibility, then the account is narrower than we might have thought, and as we noted above, it then requires a further framework for accounting for blameworthiness and praiseworthiness. If so, the Triton approach has an advantage in connecting degrees of control or quality of opportunity not only to responsible agency, but also directly to praiseworthiness and blameworthiness.

Alternatively, it might be replied that one's global view of oneself (or mechanisms) as one's own, or the appropriate targets of blame and praise, overrides any local disavowal. First, it is not clear why the overriding should go in this direction rather than the opposite. Second, an even more serious problem with this approach is that, again, because of the importance of the ecological context, it would

seem *inaccurate* to take a global view about all that issues from either oneself or from select mechanisms. For example, while it might be accurate to see oneself as an apt target of blame or praise for putting coins in the coffee tin or in an unreflective way giving a reassuring wave to a visibly uncertain driver, it would seem inaccurate to see oneself in this way when it comes to actions done in cases of severe duress. To sort out which things to take responsibility for would either presuppose an already more complete account of responsible agency that incorporates situational factors, or, at the least, would require much more in the way of explanation than moderate reasons-responsiveness currently provides.

There is a related challenge for the Taking Responsibility condition, specifically in connection with providing a role for reflection. While taking responsibility as Fischer conceives it is undoubtedly an important phenomenon, it is not clear that it can do all the work initially expected of a reflection condition. Though it captures an ability to gain reflective distance of a sort, by itself it does not entail an ability to step back and evaluate individual beliefs and desires. This seemed to be at least part of what seemed important, not only for distinguishing persons from other animals, but also for distinguishing between free and responsible agents from others.

For these reasons, among others (see, e.g., Vargas 2013, chapter 9), we believe that it is worth exploring alternative ways to accommodate the insight that reflective capacity seems to be a distinctive feature of responsible beings. There are at least three prominent roles for reflection to play in responsible agency, all compatible with each other and none requiring an additional necessary condition on responsible agency.

The first role for a capacity for reflection is embedded in the capacity to recognize certain *kinds* of reasons, namely, moral ones. To even have a conceptual grasp of other people's rights and interests, one must *have* the concept of their having desires and be able to prioritize the promotion of some over others. This ability to make second-order judgments is itself a kind of reflective capacity, and some exercise of it is required even to acquire the relevant moral concepts. In turn, this is necessary so that one can then be in a position to be responsive to morally salient features of situations even when one is not actually reflecting. This is, admittedly, a relatively thin notion of reflection. Yet it is already robust enough to mark the distinction between persons and non-persons (see Brink (2021), p. 55 and Nelkin (2020), pp. 301-302). This is a resource that is in theory available to all reasons-responsiveness accounts without adding any extra components.

For all the work reflection does in enabling us to grasp moral reasons, this does not exhaust its importance in our lives. One might think we therefore need to identify a "thicker" role for reflection. In particular, we might also think that deliberation–or the capacity to deliberate–also plays a prominent role in responsible agency. Even if it is relatively rare, deliberative activity is often taken to be the paradigm of responsible agency. It is this sort of activity that Korsgaard suggests is a key point of our being able to take "reflective distance" on our own desires and beliefs. Here things are trickier. If we can be perfectly reasons-responsive, even to moral reasons, without deliberation, and being responsive to

such reasons is sufficient for moral responsibility, blameworthiness and praiseworthiness on a given occasion, why is our having the general capacity necessary?

Here we realize that what we say is more controversial. We reject the idea that deliberative capacity is strictly entailed by reasons-responsiveness, but we accept (1) that it is necessary given human fallibility to exercise it on certain occasions, and (2) that it offers a kind of guiding ideal. Here we sketch these two basic ideas, which we have developed elsewhere in more detail.[9]

Because we are fallible and the world is not perfectly congenial, the only systematic way that we have to correct ourselves is through deliberation. Having the opportunity to step back and assess our own first-level attitudes is the one way we can do this. While it is inefficient to do this on all occasions, the capacity to do so, and our actually taking the opportunity in important moments, is what allows us to be the kind of planning and generally reasons-responsive creatures that we are. Further, deliberating—sometimes, and especially when we go wrong—can make us *more* reasons-responsive than we would otherwise be. This means that deliberative agency functions as a kind of "normative ideal," something worth seeking to realize to the extent to which one can.

The moves we have been considering seem to us open to Fischer as well. What is important for our purposes here is that they are not already incorporated, even by the Taking Responsibility condition that initially seemed the best place to capture the importance of reflection. Neither of these roles for deliberative capacity suggests that it is strictly speaking necessary for reasons-responsiveness, and we realize that this might not fully vindicate the initial thought about its essential role. But we believe that they capture the *importance* of deliberative capacity for responsible agency, if not its necessity, and that this is actually what is called for by the facts on the ground.

5. Conclusion

It will come as no surprise that we're inclined to think that Triton theories fare especially well on the three issues we've discussed in this chapter. Such theories seem to have further—or at least more readily available—resources for addressing challenges from the significance of circumstances, the location of rational capacities, and the role of reflection. Still, we won't pretend that any of these, or even the collection of these things, are jointly decisive about what version of reasons-responsiveness one should adopt. First, theory selection almost never turns on a specific issue or even a selective set of issues. It invariably depends on the wider balance of issues, the costs and benefits to a given approach in comparison to the costs and benefits of alternative approaches. Second, views are not static. Fischer and proponents of Fischer-style reasons-responsiveness might be willing to take on board some or all the features that Triton Theories employ to deflect the challenges we've discussed. Third, a yet further approach to reasons-responsiveness might fare better than either of the main alternatives we have

[9] See Brink (2021), pp. 64-67, and Nelkin (2020) for further discussion of the first approach, and Vargas (2018b) and Nelkin (2020) for discussion of the second.

discussed. Even so, we hope that the map we provided of some of the choice points might help others find their way through the rich resources that Fischer and others have developed for theories of responsibility.[10]

[10] We thank our colleagues and students in our 2020 graduate seminar on free will, where we first presented some of these ideas. Thanks also to conversations about some of these issues with Tim Bayne and David Brink. Most of all, we thank John Fischer for showing us how to do philosophy with generosity, imagination, and care.

Works cited

Arpaly, N., & Schroeder, T. (2014). *In Praise of Desire*. Oxford University Press.

Arpaly, N. & Schroeder, T. (2012). Deliberation and Acting for Reasons. *The Philosophical Review* 121, 209–39.

Bateson, M., Nettle, D., &Roberts, G. (2006). Cues of Being Watched Enhance Cooperation in a Real-World Setting. *Biology Letters* 2, 412-414.

Bierria, A. (2014). Missing in Action: Violence, Power, and Discerning Agency. *Hypatia*, 29(1), 129-145.

Bratman, M. E. (2000). Reflection, Planning, and Temporally Extended Agency. *The Philosophical Review*, *109*(1), 35-61.

Brink, D. (2021). *Fair Opportunity and Responsibility*. Oxford, U.K.: Oxford University Press.

Brink, D. O., & Nelkin, D. (2013). Fairness and the Architecture of Responsibility. *Oxford Studies in Agency and Responsibility*, *1*, 284-314.

Coates, D. J., & Swenson, P. (2013). Reasons-responsiveness and degrees of responsibility. *Philosophical Studies*, 165, 629-645.

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

Fischer, J.M. (2018). On John Doris' *Talking to Ourselves*. *Social Theory and Practice* 44, 247-253.

Fischer, J. M. (2012). *Deep Control: Essays on Free Will and Value*. Oxford: Oxford University Press.

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

Fischer, J. M., Kane, R., Pereboom, D., & Vargas, M. (forthcoming). *Four Views on Free Will* (2nd ed.). Malden, MA: Wiley-Blackwell.

Fischer, J. M., & Tognazzini, N. (2011). The Physiognomy of Responsibility. *Philosophy and Phenomenological Research*, *82*(2), 381-417.

Frankfurt, H. (1971). Freedom of the Will and the Concept of a Person. *Journal of Philosophy*, 68(1), 5-20.

Korsgaard, C. (2009). *Self-Constitution: Agency, Identity, and Integrity*. New York: Oxford University Press.

Lugones, M. (2003). Playfulness, "World"-Traveling, and Loving Perception. *Pilgrimages/Peregrinajes: Theorizing Coalition Against Multiple Oppressions* (p. xiii, 249). Lanham, Md.: Rowman & Littlefield.

Mackenzie, C. (2015). Responding to the Agency Dilemma: Autonomy, Adaptive Preferences, and Internalized Oppression. In M. Oshana (Ed.), *Personal Autonomy and Social Oppression: Philosophical Perspectives* (pp. 48-67). New York: Routledge.

Mackenzie, C., & Stoljar, N. (Eds.). (2000). *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*. New York: Oxford University Press.

McKenna, M., & van Schoelandt, C. (2015). Crossing a Mesh Theory with a Reasons-Responsive Theory: Unholy Spawn of an Impending Apocalypse or Love Child of a New Dawn. In A. Buckareff, C. Moya, & S. Rosell (Eds.), *Agency, Freedom, and Moral Responsibility* (pp. 44-64). London, U.K.: Palgrave MacMillan.

McKenna, M. (2022). Reasons-Responsiveness, Frankfurt Examples, and the Free Will Ability. In D. Nelkin and D. Pereboom (Eds.) *The Oxford Handbook of Moral Responsibility.*

Morton, J. (2011). Towards an Ecological Theory of the Norms of Practical Deliberation. *European Journal of Philosophy*, 19(4), 561-584.

Morton, J. (2022). A Moral Psychology of Poverty? In M. Vargas & J. Doris (Eds.), *The Oxford Handbook of Moral Psychology* (pp. 877-892). New York: Oxford University Press.

Nelkin, D. K. (2011). *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.

Nelkin, D. K. (2012). Responsibility and Self-Deception: A Framework. *Humana.Mente Journal of Philosophical Studies*, 20, 117–139

Nelkin, D. (2016). Difficulty and Degrees of Praiseworthiness and Blameworthiness. *Noûs*, 50(2), 356-378.

Nelkin, D. (2020). Responsibility, Reflection, and Rational Ability. *The Monist*, 103, 294–311.

Railton, P. (2009). "Practical Competence and Fluent Agency," in *Reasons for Action*. Sobel and Wall, eds., New York: Cambridge University Press, 81–115.

Schutte, O. (1998). Cultural Alterity: Cross-Cultural Communication and Feminist Theory. *Hypatia*, 13(2), 53-72.

Sripada, C. (2019). The Fallibility Paradox. *Social Philosophy and Policy*, 36(1), 234-248.

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

Vargas, M. (2018a). The Social Constitution of Agency and Responsibility: Oppression, Politics, and Moral Ecology. In M. Oshana, K. Hutchinson, & C. Mackenzie (Eds.), *The Social Dimensions of Responsibility* (pp. 110-136). New York: Oxford University Press.

Vargas, M. (2018b). Reflectivism, Skepticism, and Values. *Social Theory and Practice*, 44(2), 255-266.

Vargas, M. (2020). The Philosophy of Accidentality. *The Journal of the American Philosophical Association*, 6(4), 391-409.

Vargas, M. (2021). Constitutive Instrumentalism and the Fragility of Responsibility. *The Monist*, 104(4), 427-442.

Vargas, M. (2022). If Aristotle had Cooked: The Philosophy of Sor Juana. *Journal of Mexican Philosophy*, 1(1), 13-38.

Vargas, M. (forthcoming). Revisionism. In *Four Views on Free Will* (2nd ed., pp. tbd-tbd). Malden, MA: Wiley-Blackwell.

Vihvelin, K. (2004). Free Will Demystified: A Dispositional Account. *Philosophical Topics* 32, 427-50.

Watson, G. (1975). Free Agency. *Journal of Philosophy*, 72(8), 205-220.

Wolf, S. (1987). Sanity and the Metaphysics of Responsibility. In F. D. Schoeman (Ed.), *Free Will* (2nd ed., pp. 46-62). New York: Cambridge University Press.

Wolf, S. (1990). *Freedom Within Reason*. Oxford: Oxford University Press.