Reasons-Responsiveness Theories and the Fallibility Paradox

Word count: **2817**

Abstract: In his recent article, "The Fallibility Paradox," Chandra Sripada argues that *reasons-responsiveness* theories must hold agents morally responsible for certain rare errors that they are powerless to prevent. Moreover, he contends that *valuationist* views can avoid this paradox, because they can make fine-grained distinctions where reasons-responsiveness theories cannot. I dispute both claims. First, reasons-responsiveness theories can make the appropriate fine-grained distinction by grounding moral responsibility in two scalar properties: (1) *capacities* for reasons-responsiveness, and (2) a fair *opportunity* to exercise these capacities. Second, this kind of multivariate reasons-responsiveness theory has both *extensional* and *explanatory* advantages over the valuationist approach. Ultimately, I conclude that our intuitions about these cases of rare errors are tracking agential capacities and situational factors, rather than an agent's evaluative point of view.

As inherently fallible beings, we often make mistakes, despite our best efforts and intentions. Unfortunately, such *slips* can sometimes have rather harmful consequences.[1] A prominent line of criticism towards *reasons-responsiveness* theories of moral responsibility is that they cannot capture intuitions regarding these slips (or performance errors/mistakes).[2] Broadly speaking, reasons-responsiveness theories ground moral responsibility in agents' *capacities* to recognize and respond to reasons for action.[3] Critics have mostly argued that these theories are *under*-inclusive with respect to slips – as agents seem morally responsible for certain slips, despite lacking these capacities. Reasons-responsiveness theorists have disputed this claim in various ways, and the debate has mostly revolved around this charge.[4] Recently, however, Chandra Sripada (2019) argues that reasons-responsiveness theories are *over*-inclusive with respect to certain slips. Moreover, he contends that these slip cases

[1] I follow Amaya (2013, 2015) in defining slips as intentional actions that do not correspond to what the agent preferred to do at that time.
[2] For example, see Doris (2002), McKenna (2008), and Smith (2005, 2008).
[3] Reasons-responsiveness theories include Fischer & Ravizza (1998), Nelkin (2011), Brink & Nelkin (2013), and Sartorio (2016).
[4] For example, see Brink (2013), Amaya & Doris (2015), Murray & Vargas (2020)

constitute a strong argument in favor of a rival theory of moral responsibility, which he calls *valuationist*.[5]

In the following paper, I defend reasons-responsiveness theories against Sripada's argument, and explain why his slip cases actually support reasons-responsiveness theories. In Section 1, I present Sripada's central slip case, which involves an agent performing an iterated psychological task that produces rare but inevitable errors. Sripada argues that this kind of slip case results in an inconsistency for reasons-responsiveness theories that he terms the "fallibility paradox." In Section 2, I deny that reasons-responsiveness theories are forced into this paradox. Specially, I argue that the fallibility paradox only arises if reasons-responsiveness theories are committed to the claim that capacities of reasons-responsiveness are *sufficient* for moral responsibility. Yet, any plausible reasons-responsiveness theory holds that these capacities are merely *necessary*, along with an adequate *opportunity* to exercise these capacities. Finally, in Section 3, I argue that reasons-responsiveness theories actually have both *extensional* and *explanatory* advantages over valuationist views when it comes to Sripada's slip cases.

1. The Fallibility Paradox

Sripada's counterexample to reasons-responsiveness theories is based on a classic tool of psychological studies – the *Stroop task*. In each trial of an iterated Stroop task, the subject is shown a color word (e.g. 'red') printed in a particular color of ink. The subject is then asked

[5] Sripada (2016) independently argues for a version of this theory. In general, valuationist theories are in the tradition of *real-self/deep-self* views; for example, Frankfurt (1971) and Watson (1976).

to report the *ink* color, rather than the word. Across trials, these two things – the color word and ink color – sometimes conflict. As word reading is the default response, subjects must then use top-down regulatory processes to bias their response in favor of color naming. Most of the time, subjects effectively exercise this kind of top-down control. Still, given enough trials, there are always errors.[6] Some of these errors are attributable to lapses in concentration, or other factors that undermine top-down control. However, Sripada claims that at least some of these errors are due to the probabilistic nature of this top-down control *itself*. In other words, although subjects can exercise control to *bias* their response, sometimes the wrong response is still produced. These "noise-based" errors (or slips) result from the fact that, "human information processing involves an ineliminable role for stochasticity." (p. 241)

Sripada asks us to imagine the case of Fei, an undergraduate student taking part in a psychological experiment utilizing the Stroop task:

> The task involves 1,000 trials over the course of about forty minutes. Her performance on the task is incentivized: for each trial with a correct response, twenty cents is donated to the local branch of the Humane Society, a charity that helps stray animals in the community. For each incorrect response, twenty cents is deducted from the amount to be donated. Fei cares very much about animals – over the years, she has fostered several animals without homes –and so she tries very hard, and equally hard, on every single trial of the task. She produces the correct response on 996 trials and makes just four [noise-based] errors. (p. 235)

Sripada concludes from this case that Fei possess the appropriate capacities for reasons-responsiveness. Although there are different explications of reasons-responsiveness, Fei's high degree of accuracy at the task suggests that she satisfies any reasonable conception.

---

[6] Sripada (2019) reports that standard error rates are 5-10% in trials where the color word and ink color conflict.

Sripada then claims that Fei is not morally responsible for her four (noise-based) errors. In particular he appeals to the principle that, "a person cannot be morally responsible for something that she doesn't want to happen and is powerless to prevent." (p. 244) Given these two plausible claims, Sripada then argues that reasons-responsiveness theories are forced into inconsistency. After all, since Fei is reasons-responsive, Sripada infers that she must be morally responsible for each trial, *including* for her errors. Yet, surely reasons-responsiveness theories want to capture the intuition that Fei is *not* morally responsible, based on the plausible principle. Sripada calls this kind of inconsistency the *fallibility paradox*. Moreover, he claims that reasons-responsiveness theories will encounter this paradox – that an agent is morally responsible for rare errors, despite being powerless to prevent them – in any relevantly similar situation.[7]

Sripada's explanation for why reasons-responsiveness theories are vulnerable to the fallibility paradox is "[their] use of seemingly arbitrary thresholds…" (p. 238) In other words, these theories must set some threshold for reasons-responsiveness.[8] Usually, these thresholds are explained by appealing to counterfactuals in terms of possible worlds. In this way, an agent is reasons-responsive to the degree that, in relevant possible worlds where there is sufficient reason for her to act a certain way, she recognizes this reason and responds

---

[7] In particular, the fallibility paradox seems to generalize to any situation where the agent performs an iterated psychological task that produces rare but inevitable errors, despite the fact that she is deeply committed to doing her best.

[8] Reasons-responsiveness is normally a composite of capacities to *recognize* (cognitive) and *respond* (motivational) to reasons. On some accounts, such as Fischer & Ravizza (1998), these two capacities can have asymmetric thresholds. However, in order to simplify, I will assume that they have similar thresholds. As will become clear, nothing in my argument hangs on this issue.

accordingly.[9] The higher the required degree of response, the higher the threshold. Yet, even if this threshold is set fairly high, agents like Fei will clear it. After all, Fei is *remarkably* reasons-responsive when it comes to the Stroop task – she recognizes and responds to the relevant reasons over 99% of the time. Yet, because Fei clears the threshold for this task, she is morally responsible for *all* trials, if reasons-responsiveness is sufficient for responsibility. It is this upshot that leads to the fallibility paradox.

2. Response to the Fallibility Paradox

In characterizing reasons-responsiveness theories, Sripada primarily draws on the work of John M. Fischer & Mark Ravizza (1998). However, reasons-responsiveness theories need not share all the features of this particular view. For instance, reasons-responsiveness theories are not committed to thresholds. Indeed, given that reasons-responsiveness is usually explained in terms of a proportion of possible worlds, this property could easily be *scalar*.[10] Of course, this would mean that moral responsibility would also be a scalar property, but there are already independent reasons for this conception.[11] Now, going scalar does not avoid the fallibility paradox. After all, Fei is *highly* reasons-responsive; and so, she would be highly morally responsible, according to the argument. Still, it is important to accurately diagnose the source of the fallibility paradox, and the use of thresholds is not it. Instead, the source of the paradox is the fact that reasons-responsiveness theories must seemingly treat

---

[9] Throughout the paper, including here, I assume an *agent*-based version of a reasons-responsiveness theory, based on Brink & Nelkin (2013). Sripada claims that the fallibility paradox applies to both agent-based and mechanism-based views. Thus, I use an agent-based view for ease of exposition.

[10] Indeed, Brink & Nelkin (2013) suggest such a scalar view.

[11] One independent reasons for a scalar conception of moral responsibility is that the excuses of *coercion* and *duress* admit of degree, and so they presumably diminish moral responsibility in degrees.

all instances of an iterated task the same[12]. In other words, what matters for reasons-responsiveness is the agent's *overall* responsiveness to the task. This level of responsiveness then applies to *every* instance. Yet, our intuition in cases like Fei is that some instances should be treated differently (i.e. noise-based errors).

Regardless of thresholds, then, reasons-responsiveness theories face the fallibility paradox if moral responsibility is simply a product of an agent's capacity for reasons-responsiveness.[13] Again, though, reasons-responsiveness theories are not committed to this view. Instead, many reasons-responsiveness theories actually hold that moral responsibility for wrongdoing is a product of *both* an agent's capacity for reasons-responsiveness, and a fair (or adequate) *opportunity* to exercise this capacity. There are different ways of explicating these two factors, but the underlying idea is that responsibility for wrongdoing requires certain (internal) conditions of the *agent*, and certain (external) conditions of her *situation*. In this way, even if an agent is reasons-responsive, her situation must also provide her a fair opportunity to avoid wrongdoing.[14]

Given this multivariate conception of a reasons-responsiveness theory, the fallibility paradox can be defused. After all, consider the four trials that Fei failed. The common feature of these trials is that *noise* caused the error. Specifically, the inherent stochasticity of Fei's top-down control produced a response that ran contrary to her intentions. Yet, if Fei was *truly* powerless to prevent this error, then it seems that her situation did not provide a

---

[12] Indeed, towards the end of the paper, Sripada (2019) characterizes the fallibility paradox as a "problem of coarse-grainedness." (p. 247). I believe that this is a better explanation of the source of the inconsistency for reasons-responsiveness theories.

[13] Like Sripada, I am assuming that all other conditions for moral responsibility are fulfilled.

[14] I draw on Brink & Nelkin (2013) here.

fair opportunity in these trials; and thus, she is excused on these grounds. Contrast this error with one that does not involve powerlessness. For instance, suppose that Fei fails a trial simply because she gets bored with the task and stops paying attention. In this case, she retains the fair opportunity to avoid wrongdoing, and so she is still morally responsible. In this way, reasons-responsiveness theories can vindicate the principle that a person cannot be morally responsible for something that she doesn't want to happen and is powerless to prevent – avoiding the paradox.

Another way to state this response to the fallibility paradox is in the form of a dilemma. In particular, the reasons-responsiveness theorist can ask Sripada to describe the trials of the given task that purportedly excuse. If he provides trials that violate the principle of powerlessness, then the source of this powerlessness is a plausible basis for a lack of fair opportunity; and the agent is excused on the proposed reasons-responsiveness account. On the other hand, if Sripada provides trials that do *not* violate the principle, then the initial motivation for excuse loses its force. Now, one might argue that the force of the paradox can be preserved in cases where, although not *entirely* powerless, the agent still finds it extremely *difficult* to avoid erring. For instance, we might think that maintaining full attention throughout a task with many trials is sufficiently difficult, such that if an agent fails a trial due to a lapse in concentration, she ought to be excused. However, a reasons-responsiveness theory can even account for this intuition. After all, just as reasons-responsiveness can be scalar, so too can the notion of fair opportunity. Thus, in trials where concentration lapses due to difficulty in maintaining attention, this difficulty is a plausible basis for *reduced*

opportunity – and the agent's moral responsibility is proportionally diminished. In this way, the whole force of the fallibility paradox is neutralized.

3. Comparison to Valuationist Views

Reasons-responsiveness theories can avoid the fallibility paradox primarily because reasons-responsiveness is not a *sufficient* condition for moral responsibility, it is merely a *necessary* one. However, it is worth considering whether it is the *best* theory of responsibility for the kind of errors at hand. Sripada himself argues that reasons-responsiveness theories are inferior to so-called *valuationist* views.[15] According to this family of views, "for an agent to be morally responsible for an action, the action must flow from and express her evaluative point of view." (p. 246) As valuationist views are a significant rival to reasons-responsiveness theories, comparing the two on this issue is valuable to larger debates regarding moral responsibility. To this end, I will compare the two theories in terms of *extensional adequacy* and *explanatory power*; that is, I will consider which theory better captures intuitions, and provides the better explanation of moral responsibility, in these cases?

In Section 2, I argued that reasons-responsiveness theories are extensionally adequate in the relevant cases of error. In fact, not only can they capture bivalent intuitions about moral responsibility, but they also have the (scalar) resources to capture more nuanced intuitions. What about valuationist views? Sripada claims that they make the right predictions in the same cases. For example, in the four trials that Fei failed, he argues that she is excused on a valuationist view because, "her actions don't flow from her values." (p. 247) This seems

[15] As I mentioned in f. 5, valuationist views are in the tradition of real-self/deep-self theories, like Frankfurt (1971) and Watson (1976). However, it seems plausible that the label also includes some *quality-of-will* views.

plausible, although I wonder if this view has the resources to intuitively capture cases of partial responsibility. Setting aside the issue of partial responsibility, though, valuationist views get things wrong in other cases. For instance, suppose that Wei is another undergraduate student taking part in the psychological experiment. Although she also cares very much about animals, she is much lazier than Fei; and so, she does not try as hard on the Stroop task, and ends up with many more errors. Now, assume that some of these errors are *not* noise-based. The valuationist view seems forced to conclude that Wei is not morally responsible for *these* errors either. After all, as an animal-lover, these errors do not flow from her values. Yet, this contradicts our intuition, as Wei's mistakes are not innocent (or inevitable) in the same way as Fei. In contrast, reasons-responsiveness theories can straightforwardly capture this intuition, since Wei is reasons-responsive and had a fair opportunity to avoid wrongdoing in these trials.

Now, Sripada intended for the fallibility paradox to highlight errors that agents are *not* morally responsible for committing. Thus, the case of Wei is technically outside the boundaries of this specific issue. Are valuationist views at least extensionally adequate within this domain? I argue that they are not. Indeed, consider Donna – another student involved in the experiment. Unlike both Fei and Wei, Donna does not care about animals. However, she recognizes that other people do, and so she makes an honest attempt to get every trial of the Stroop task correct. Now, suppose that Donna also produces four noise-based errors. It seems plausible that Donna is morally responsible on a valuationist view, since these failures are expressive of her evaluative point of view (she does not care about animals). Yet,

intuition suggests that *regardless* of her evaluative point of view, Donna should be excused for her noise-based errors. Thus, valuationist views get things wrong in this case, as well.

Sripada would likely respond that these noise-based errors do not *flow* from Donna's values, since the cause of the errors is not part of her evaluative point of view. Indeed, Sripada's own valuationist account holds that, "an action expresses an element of one's evaluative point of view if that element motivationally supports performing the action." (p. 246, f. 21) On this kind of *causal* account, noise-based errors are seemingly excusing by definition, since they are caused by the inherent stochasticity of certain informational processing. Still, I believe that even this casual account is vulnerable to counterexample. Imagine that Donna produces *another* error because it is extremely difficult for her to concentrate due to a recent personal crisis. Moreover, suppose that if Donna *did* care about animals, this would provide a motivational boost that would allow her to overcome this difficulty. In this case, it is more plausible that the error expresses her evaluative point of view, as her lack of care motivationally supports her failure (counterfactually). Yet, Donna does not seem particularly morally responsible for this error; and so, valuationist views get things wrong again. Reasons-responsiveness theories, on the other hand, can easily capture this intuition by appealing to her lack of a fair opportunity, due to difficulty. In this way, reasons-responsiveness theories are more extensionally adequate than valuationist view.

Furthermore, reasons-responsiveness theories provide a better *explanation* of moral responsibility in these cases than valuationist views. Recall that Sripada's own principle to motivate the intuition that Fei is not morally responsible is that a person cannot be morally responsible for something that she doesn't want to happen and is powerless to prevent.

Insofar as this is a plausible principle, it is really the powerlessness of the agent that seemingly drives intuitions regarding noise-based errors. This explains why Donna is intuitively excused for her noise-based errors, even though she does not care about animals. But this notion of powerlessness is much more consistent with reasons-responsiveness theories. In particular, it is especially consonant with the (external) situational element of a multivariate account, which captures the idea that agents should have a fair opportunity to avoid wrongdoing. For valuationist views, on the other hand, powerlessness is only significant insofar as it indirectly influences the expression of one's evaluative point of view. Yet, powerlessness seems *directly* morally relevant, regardless of its relation to our values.

Now, one could dispute that powerlessness is mostly driving intuitions in these cases. After all, Fei cares about animals, and even Donna wants to get things right. Thus, it could be that the "doesn't want to happen" component of the principle is also significantly responsible for intuitions. Moreover, it might seem that this component is more amenable to a valuationist gloss. Even if we factor in this component, though, it does not give valuationist views any explanatory advantage. First of all, suppose that we stipulate that Donna does *not* want to get the trials right. Still, I contend that our intuitions regarding her moral responsibility for noise-based errors do not shift – she is still plausibly excused. This suggests that it is indeed her powerlessness that largely motivates our intuitions. Secondly, even supposing that Donna does want to get things right – in *some* sense – this motivational element of her psyche is not meaningfully part of her evaluative point of view. Thus, insofar

as this other component of the principle might subtly affect our intuitions, it neither requires nor recommends a valuationist explanation.[16]

[16] Indeed, if valuationist views tried to capture such cases, by claiming that even these kinds of weak wants/desires can constitute part of one's evaluative point of view, this would lead to a rather deflationary theory that is implausible.

Bibliography

Amaya, S. (2013). Slips. *Noûs*, 47 (3), 559–576.

Amaya, S., & Doris, J. (2015). No excuses: Performance mistakes in morality. In J. Clausen and N. Levy (Eds.), *Handbook of neuroethics*, (pp. 253–272). Dordrecht: Springer.

Brink, D., & Nelkin, D. (2013). Fairness and the architecture of responsibility. In D. Shoemaker (Ed.), *Oxford studies in agency and responsibility* (Vol. 1, pp. 284-313). Oxford: Oxford University Press.

Brink, D. (2013). Situationism, responsibility, and fair opportunity. *Social Philosophy and Policy*, 30 (1–2), 121–149.

Doris, J. M. (2002). *Lack of character: Personality and moral behavior*. New York: Cambridge University Press.

Fischer, J., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy*, 68 (1), 5–20

McKenna, M. (2008). Putting the lie on the control condition for moral responsibility. *Philosophical Studies,* 139 (1), 29-37.

Murray, S., & Vargas, M. (2020). Vigilance and control. *Philosophical Studies*, 177 (3), 825-843.

Nelkin, D. (2011). *Making sense of freedom and responsibility*. Oxford: Oxford University Press.

Rudy-Hiller, F. (2019). Give people a break: Slips and moral responsibility. *Philosophical Quarterly*, 69 (277), 721-740.

Sartorio, C. (2016). *Causation and free will*. Oxford: Oxford University Press.

Smith, A. (2005). Responsibility for attitudes: Activity and passivity in mental life. *Ethics*, 115 (2), 236–271.

Smith, A. (2008). Control, responsibility, and moral assessment. *Philosophical Studies*, 138, 367–392.

Sripada, C. (2016). Self-expression: A deep self theory of moral responsibility. *Philosophical Studies*, 173, 1203–1232.

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

Watson, G. (1975). Free agency. *The Journal of Philosophy*, 72 (8), 205–220.